UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Métodos de estimação baseados em modelos na presença de dados faltantes

Taís Roberta Ribeiro

Tese de Doutorado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP SÃO CARLOS

Taís Roberta Ribeiro

Métodos de estimação baseados em modelos na presença de dados faltantes

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Doutora em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientadora: Profa. Dra. Daiane Aparecida Zuanetti

USP – São Carlos
Novembro de 2022

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

R484m

Ribeiro, Taís Roberta
Métodos de estimação baseados em modelos na presença de dados faltantes / Taís Roberta Ribeiro; orientador Daiane Aparecida Zuanetti. -- São Carlos, 2022.
178 p.

Tese (Doutorado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2022.

1. modelos lineares e não lineares de regressão. 2. imputação de dados . 3. integração numérica. 4. algoritmo EM. I. Zuanetti, Daiane Aparecida , orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Taís Roberta Ribeiro

Model-based estimation methods in the presence of missing data

Doctoral dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar, in partial fulfillment of the requirements for the degree of the Doctorate Interagency Program Graduate in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Profa. Dra. Daiane Aparecida Zuanetti

USP – São Carlos
November 2022

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Defesa de Tese de Doutorado da candidata Taís Roberta Ribeiro, realizada em 14/10/2022.

Comissão Julgadora:

Profa. Dra. Daiane Aparecida Zuanetti (UFSCar)

Prof. Dr. Luis Aparecido Milan (UFSCar)

Profa. Dra. Teresa Cristina Martins Dias (UFSCar)

Prof. Dr. Erlandson Ferreira Saraiva (UFMS)

Profa. Dra. Lia Hanna Martins Morita (UFMT)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa Interinstitucional de Pós-Graduação em Estatística.

AGRADECIMENTOS

Primeiramente, gostaria de agradecer a Deus e a todos da minha família, em especial meus pais e irmãos, pela força diária em cada momento passado durante meu doutorado, por sempre estarem ao meu lado em todas as situações e por não me permitirem desistir por mais conturbadas que fossem as intempéries pelas quais passei durante esta trajetória e que, por fim, me trouxeram até aqui e me fizeram ver que tudo valeu a pena e que foi necessário para o desenvolvimento desta pesquisa. Obrigada por me proporcionarem todo o apoio que precisei para a realização deste sonho! Agradeço também à minha sobrinha Helena, que em tão poucos meses já fez tanto por mim e se tornou um dos motivos para eu sempre buscar minha melhor versão em todas as esferas da minha vida. Agradeço ao meu cunhado Renato pelos conselhos acadêmicos e por sempre se disponibilizar a ajudar.

Gostaria de agradecer aos meus amigos que permanecem comigo desde a época da minha graduação, alguns infelizmente não conseguimos nos ver tanto quanto gostaríamos, mas sei que sempre estão torcendo por mim, Alex, Amabele, Danielle, Donizetti, Félix, Gabrielle, Liara e Marina. Um agradecimento especial ao Alex por emprestar seu computador para realizar algumas simulações necessárias para esta pesquisa e por sempre se preocupar comigo e melhorar meus dias com nossos divertimentos culturais que tanto gostamos, como teatros, cinemas, filmes do Oscar, entre outros. Obrigada Alex, Diego e Félix por estes passeios, foram muito importantes para mim e que aproveitemos muito mais juntos ainda.

Quero agradecer também aos amigos e pessoas iluminadas que conheci em São Carlos, por cada conselho, cada palavra de conforto dada, por me ouvirem e também por todas as histórias vividas juntos e que me trouxeram a alegria que eu precisava em momentos mais difíceis, Alana, Bianca, Franciele, Isabella, Letícia, Luciana, Tainá, Talia e Tiago. Obrigada por acreditarem em mim e também por cada conquista que sempre comemoramos juntos.

Um agradecimento também à duas pessoas especiais que cruzaram meu caminho em pouco tempo que me mudei para São Paulo. Obrigada Beatriz, a melhor bailarina do mundo, pela sua energia incrível, estar ao seu lado é sempre ter a sensação que o melhor nos espera, obrigada por me ensinar a sempre ver o lado bom das coisas. Obrigada Juliana, pelas palavras de motivação na reta final do meu doutorado, quando eu estava com várias inseguranças você foi uma das pessoas que me ajudaram a lidar com elas e me fizeram enxergar que tudo daria certo.

Por fim, agradeço à CAPES pelo auxílio financeiro durante quase quatro anos para que esta pesquisa fosse realizada, aos professores que compuseram a minha banca de defesa desta tese e, em especial, à minha orientadora Daiane e à professora Teresa Cristina do Departamento

de Estatística da Universidade Federal de São Carlos (UFSCar), elas sabem o quanto sou extremamente grata por estarem comigo durante o desenvolvimento desta pesquisa e por toda a aprendizagem que me proporcionaram.

RESUMO

RIBEIRO, T. R. **Métodos de estimação baseados em modelos na presença de dados faltantes**. 2022. 178 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2022.

Os dados faltantes são observações que deveriam ter sido feitas, mas não foram por algum motivo, reduzindo, assim, a capacidade de entender a natureza do fenômeno, além de dificultar a extração de informações através dos dados analisados, já que o impacto nos resultados dos estudos nem sempre são conhecidos. Como uma considerável parte das técnicas estatísticas foram desenvolvidas para analisar dados completos, os dados faltantes geralmente precisam ser tratados de maneira que o conjunto de dados resultante possa ser analisado por tais métodos já consolidados. Os métodos mais utilizados para lidar com dados faltantes se dividem, principalmente, entre métodos de remoção e de imputação de dados, sendo ambas as configurações, na maioria das vezes, desvantajosas em termos da análise do resultado final, seja por tornar os resultados viesados ou por termos que trabalhar com a incerteza associada à imputação de valores desconhecidos. Nesse trabalho, então, propomos alguns métodos baseados em modelos para a resolução do problema de dados ausentes para análise de regressão, sem que seja necessário recorrer à imputação ou à remoção de informações. Verificamos o desempenho das metodologias propostas em dados simulados sob diferentes cenários e comparamos com o desempenho de outras técnicas tradicionais de imputação e remoção de dados.

Palavras-chave: modelos lineares e não lineares de regressão, imputação de dados, integração numérica, algoritmo EM.

ABSTRACT

RIBEIRO, T. R. **Model-based estimation methods in the presence of missing data**. 2022. 178 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2022.

The missing data are observations that should have been made, but were not for some reason, thus reducing the ability to understand the nature of the phenomenon, in addition to making it difficult to extract information from the analyzed data, since the impact on the results of the studies is not always known. As a considerable part of the statistical techniques were developed to analyze complete data, the missing data usually need to be treated in such a way that the resulting dataset can be analyzed by such established methods. The most used methods to deal with missing data are divided, mainly, between methods of data removal and imputation, being both configurations, in most cases, disadvantageous in terms of the analysis of the final result, either by making the results biased or because we have to work with the uncertainty associated with the imputation of unknown values. In this work, then, we propose some model-based methods for solving the problem of missing data for regression analysis, without having to resort to imputation or removal of information. We verified the performance of the proposed methodologies on simulated data under different scenarios and compared it with the performance of other traditional techniques of imputation and data removal.

Keywords: linear and nonlinear regression models, data imputation, numerical integration, EM algorithm.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Nos dias de hoje, devido à facilidade na automação de processos, à alta capacidade de armazenamento das mídias e ao uso generalizado de computadores, muitos dados estão sendo coletados (BROWN; KROS, 2003). Profissionais e pesquisadores reconhecem que vivemos na "era do *big data*", havendo até mesmo os que falam em *zettabytes* (KURASOVA *et al.*, 2014; TAN; TSANG; WANG, 2014; WANG *et al.*, 2014). Mediante este cenário, muitas empresas estão repensando seus negócios, já que, para melhorar seu desempenho e, consequentemente, obter vantagens competitivas no mercado, é essencial conseguir coletar e analisar seus dados de forma consistente (RIBEIRO, 2015). Atualmente, a informação conseguiu adquirir um valor que há poucos anos era inimaginável, tanto é que, em muitos casos, o maior bem que a organização tem é aquilo que ela sabe sobre seus clientes (PRASS, 2014).

Porém, acompanhado com o crescimento notório de dados disponíveis originados de inúmeras fontes, devemos considerar que a garantia da qualidade destes dados também é de imensa importância para a elaboração de um estudo (WU; WUN; CHOU, 2004). No entanto, uma parte considerável dos bancos de dados existentes é caracterizada pela imprecisão e pela incompletude, isto é, pela presença de valores ruidosos (erros e *outliers*) e faltantes, respectivamente (FARHANGFAR; KURGAN; PEDRYCZ, 2004). Os dados faltantes (*missing data*), enfoque principal deste trabalho, são informações incompletas ou perdidas que ocorrem porque o respondente se recusa ou é incapaz de dar a resposta correta a um ou vários itens, por exemplo. Outra causa possível de não resposta está na falha do entrevistador ao perguntar ou registrar a resposta do indivíduo participante da pesquisa, fazendo com que uma possível informação seja considerada incorreta na etapa posterior de edição dos dados e análise de consistência. Além dos fatores supracitados, temos também a entrada de dados de forma manual (LAKSHMINARAYAN; HARP; SAMAD, 1999; FARHANGFAR; KURGAN; PEDRYCZ, 2004; FARHANGFAR; KURGAN; PEDRYCZ, 2007), as medições identificadas como incorretas (FARHANGFAR; KURGAN; PEDRYCZ, 2004; FARHANGFAR; KURGAN; PEDRYCZ, 2007), os equipamentos com falhas operacionais (WU; WUN; CHOU, 2004; FARHANGFAR; KURGAN; PEDRYCZ,

2004; FARHANGFAR; KURGAN; PEDRYCZ, 2007; BUUREN; MULLIGEN; BRAND, 1994; COLANTONIO *et al.*, 2010) e o alto custo de coleta de dados (MYRTVEIT; STENSRUD; OLSSON, 2001) como outras causas do surgimento de dados faltantes.

Na área de aprendizado de máquina e análise estatística de dados, nota-se uma dificuldade na fase de aprendizagem, inferência e previsão mediante a presença de dados faltantes (MARLIN, 2008). Com isso, podemos constatar que profissionais das mais diversas áreas têm cada vez mais consciência dos problemas que podem ser acarretados pelos dados faltantes (ASSUNÇÃO, 2012). Dentre as principais adversidades advindas pelos *missing*, podemos citar as complicações na manipulação e análise dos dados, a perda de eficiência e o viés, resultantes das discrepâncias entre os valores atribuídos aos dados faltantes e os valores reais desconhecidos (FARHANGFAR; KURGAN; PEDRYCZ, 2007). Como exemplo destes problemas em situações cotidianas, temos que a perda de dados representa um obstáculo no planejamento e análise dos estudos epidemiológicos, nos quais, frequentemente, a meta é determinar preditores que contribuam para prever a ausência ou presença de uma doença em uma população. Dessa forma, a inexistência de informações, tanto nos preditores como na variável resposta, pode acarretar em uma análise de dados enviesada e insatisfatória (NUNES; KLÜCK; FACHEL, 2009).

Outro exemplo de problemas na análise estatística causados por dados faltantes é observado ao analisar bases públicas com indicadores educacionais, através da Teoria de Resposta ao Item (TRI) - ramo da estatística direcionado predominantemente ao estudo de questionários e outras listas de itens - em que a utilização desta técnica por vezes é dificultada ou impossibilitada com a ausência de informações, já que ignorar tais incompletudes na base de dados pode criar problemas na estimação dos parâmetros (PEREIRA, 2014).

Podemos listar também os dados que são armazenados e analisados em tempo real, como previsão de séries de demanda de energia, de qualidade do ar, da vazão de água, além de séries financeiras. Nesses estudos, as séries devem estar ordenadas cronologicamente para que estejam aptas a serem estudadas sem a presença de valores faltantes entre medidas observadas (SANTANA; FILIZOLA-NETO; FREITAS, 2010; LOPES, 2007).

Portanto, em vista de todos estes aspectos negativos resultados dos *missing data*, muitas metodologias vêm sendo desenvolvidas com o objetivo de solucioná-los. No entanto, infelizmente, por falta de conhecimento, problemas computacionais e de tempo, dentre outros motivos, muitas delas não são utilizadas, dando abertura ao uso de abordagens mais simples que podem proporcionar mais prejuízos do que benefícios aos estudos (ASSUNÇÃO, 2012), como o preenchimento de um valor faltante por zero. Apenas a partir de 1987, com a publicação dos livros: *Statistical Analysis with Missing Data* (RUBIN; LITTLE, 2019) e *Multiple Imputation for Nonresponse in Surveys* (RUBIN, 1987) aliado ao mais fácil acesso à computação pelas pessoas, é que se passou a investir mais na solução de problemas com dados faltantes (GRAHAM *et al.*, 2009).

Neste novo cenário, dois primeiros artigos publicados que abordam um primeiro método

acessível para lidar com *missing data*, por meio da modelagem de equações estruturais, foram *Estimation of linear models with incomplete data* (ALLISON, 1987) e *On structural equation modeling with data that are not missing completely at random* (MUTHÉN; KAPLAN; HOLLIS, 1987). Neste mesmo ano de 1987, o artigo *The calculation of posterior distributions by data augmentation* colaborou com o desenvolvimento de softwares de imputação múltipla (TANNER; WONG, 1987).

Quanto aos métodos mais utilizados atualmente de tratamento de dados faltantes, temos que eles envolvem a substituição ou remoção dos mesmos. No processo de remoção, geralmente costuma-se eliminar os casos de incompletude, construindo o modelo apenas com os dados completos. Entretanto, com esta metodologia, os resultados obtidos se tornarão viesados se os casos restantes não representarem toda a população. Nesta mesma linha de procedimentos, podemos citar também a não inclusão no modelo das variáveis que possuem dados faltantes o que, embora não acarrete problemas de enviesamento da base de estimação ou aprendizado, pode ser determinante na obtenção de um modelo com um poder preditivo inferior ao que seria resultado caso todas as variáveis fossem testadas (ASSUNÇÃO, 2012).

Para contornar esse problema, temos, então, as técnicas estatísticas que envolvem imputação de dados faltantes, tendo por objetivo "completar" os mesmos e possibilitar a análise com todos os indivíduos e variáveis disponíveis no estudo. As primeiras técnicas de imputação desenvolvidas abrangiam métodos de relativa simplicidade de uso, tais como, substituição dos dados faltantes pela média, pela mediana, por interpolação ou até por regressão linear através da imputação única, ou seja, o dado ausente é preenchido uma única vez e então se utiliza o banco de dados completo para as análises. Entretanto, deve-se levar em consideração que os valores imputados não são valores reais, logo, é essencial trabalhar com a incerteza associada à imputação, visando validar os resultados obtidos com os dados completos. Sendo assim, para solucionar essa questão foi desenvolvida a técnica de Imputação Múltipla, IM (NUNES; KLÜCK; FACHEL, 2009).

Como uma forma de nos auxiliar e direcionar em relação a quando e como utilizar os métodos de imputação, temos que eles podem ser subdivididos quanto ao uso em problemas de séries temporais e problemas gerais. Em relação aos problemas com séries temporais, há os que envolvem dados sem tendência e sem sazonalidade (geralmente a imputação é feita pela média, mediana, moda ou por valor aleatório), dados com tendência e sem sazonalidade (imputação por interpolação linear) e dados com tendência e com sazonalidade (imputação mediante ajuste sazonal mais interpolação). Por outro lado, considerando agora os problemas gerais, devemos averiguar se as variáveis são categóricas (primeiramente tornamos os valores NA (*missings*) uma categoria e, em seguida, realizamos a imputação por imputação múltipla ou via regressão logística) ou contínuas (imputação por média, mediana, moda, regressão linear ou imputação múltipla).

Quanto aos trabalhos mais recentes publicados relativos às diferentes formas de lidar

com dados faltantes, podemos citar o artigo *Reporting the Use of Multiple Imputation for Missing Data in Higher Education Research* (MANLY; WELLS, 2015). Este trabalho recomenda práticas e relatos de IM que envolvem a descrição da natureza e estrutura de quaisquer dados ausentes, descrevem o modelo e os procedimentos de imputação, assim como alguns resultados de imputação notáveis. O *Group-sparse subspace clustering with missing data* (PIMENTEL-ALARCÓN *et al.*, 2016), apresenta dois novos métodos para agrupamento (*clustering*) de subespaço com dados perdidos: (a) agrupamento de subespaços esparsos de grupo (GSSC), que é baseado em dispersão de grupo e minimização alternada, e (b) mistura de subespaços de agrupamento (MSC), que modela cada observação como uma combinação convexa de suas projeções em todos os subespaços da união; *A Unified Approach to Measurement Error and Missing Data: Overview and Applications* (BLACKWELL; HONAKER; KING, 2017), oferece ilustrações empíricas, software de código aberto que implementa todos os métodos descritos, e um artigo complementar com detalhes técnicos e extensões para lidar com dados faltantes; *Recurrent Neural Networks for Multivariate Time Series with Missing Values* (CHE *et al.*, 2018), propõe novos modelos de aprendizagem profunda, a GRU-D, que considera dados faltantes em séries temporais.

Uma importante pesquisa já realizada envolvendo dados com informações faltantes é a tese *Análise de dados categorizados com omissão em variáveis explicativas e respostas* (POLETO, 2011), que apresenta métodos para analisar estes dados e também estudos delineados para compreender os resultados de tais análises. Este trabalho também mostra análises de sensibilidade Bayesiana e clássica para dados com respostas categorizadas sujeitas a omissão. Mostra-se que as componentes subjetivas de cada abordagem podem influenciar os resultados de forma não-trivial, independentemente do tamanho da amostra, e que, portanto, as conclusões devem ser cuidadosamente avaliadas. Especificamente, demonstra-se que distribuições a priori comumente consideradas como não-informativas ou levemente informativas podem, na verdade, ser bastante informativas para alguns parâmetros, e que a escolha do modelo superparametrizado é igualmente importante.

No trabalho *Sequentially additive nonignorable missing data modeling using auxiliary marginal information* (SADINLE; REITER, 2019) foi estudado uma classe de mecanismos de falta, chamados sequencialmente aditivos não-ignoráveis, com o objetivo de modelar dados multivariados com a não resposta ao item. Esses mecanismos permitem explicitamente que a probabilidade de não resposta de cada variável dependa do valor dessa variável, representando assim mecanismos de ausência não - ignoráveis. Esta é uma das mais recentes publicações sobre maneiras inovadoras de lidar com os problemas advindos do estudo de banco de dados com informações faltantes.

O objetivo principal desse trabalho, portanto, é propor duas estratégias que permitem a estimação de modelos lineares e não lineares de regressão linear múltipla na presença de dados faltantes. A primeira estratégia envolve um método baseado em modelo com resolução analítica

com uma ou duas variáveis com valores faltantes. Para ilustração dessa metodologia adotamos o modelo Gaussiano, mas outros modelos podem ser considerados desde que a solução analítica exista. A segunda estratégia aborda três métodos baseados em modelo sem resolução analítica e também com uma ou duas variáveis com valores faltantes, são eles: método por integração numérica, método utilizando média de log-verossimilhanças e método utilizando o algoritmo EM. Para ilustração da segunda metodologia, utilizamos o modelo Weibull, mas outros modelos podem ser assumidos, se necessário. Apesar de aqui abordamos as situações com uma ou duas variáveis faltantes, os métodos podem ser estendidos para mais variáveis faltantes e, ao contrário de grande parte dos métodos de imputação que apenas completam os dados faltantes, estimam o modelo final na presença deles e também podem ser utilizados como métodos de imputação, caso esse seja o objetivo em outros estudos.

Este trabalho esta estruturado em cinco capítulos. No Capítulo 2, apresentamos a caracterização do problema com dados faltantes, suas causas, seus principais tipos, padrão, quantidade e quais critérios devem ser levados em consideração no momento da seleção do método apropriado para lidar com suas diferentes variações. No Capítulo 3, estudamos os principais métodos já existentes para lidar com uma base de dados com informações faltantes. No Capítulo 4, apresentamos a primeira estratégia proposta nesta tese, que é um método baseado em modelos com solução analítica para problemas envolvendo base de dados com informações faltantes. Neste capítulo, consideramos o modelo Gaussiano de regressão linear com k variáveis como a distribuição destas variáveis e realizamos estudos de simulação considerando diversos cenários. Em seguida, comparamos os resultados inferenciais e preditivos da proposta com alguns dos principais métodos utilizados para lidar com dados faltantes. No Capítulo 5, apresentamos a segunda estratégia proposta nesta tese, que são métodos baseados em modelos para quando a solução analítica não está disponível. Os estudos de simulação são realizados considerando que a distribuição das variáveis não é a Gaussiana. Da mesma forma, comparamos os resultados inferenciais e preditivos dos métodos desenvolvidos neste capítulo com alguns dos principais métodos utilizados para lidar com dados faltantes. No Capítulo 6, apresentamos uma aplicação do método baseado em modelo para o conjunto de dados *Airquality*, comparando o resultado preditivo com os outros métodos também utilizados para comparação nos estudos de simulação. Enfim, no Capítulo 7, temos a conclusão e discussão dos principais resultados obtidos.

CAPÍTULO

2

OS DADOS FALTANTES

Este capítulo tem como objetivo apresentar os principais conceitos referentes aos dados faltantes, para que saibamos, assim, a maneira correta e mais eficaz de manipulá-los.

2.1 Definição

Primeiramente, considerando um conjunto de dados, um elemento é dito completo se todas as suas características (campos ou variáveis) estão preenchidas com dados apropriados (VERONEZE, 2011). Um dado faltante indica que uma característica de um elemento da pesquisa não está preenchida. Um caso incompleto é um elemento que contém dados faltantes (WU; WUN; CHOU, 2004).

Dessa forma, podemos afirmar que, na maior parte das vezes, os dados faltantes são observações que deveriam ter sido feitas, mas não foram por algum motivo. Isso reduz a capacidade de entender a natureza do fenômeno estudado, além de dificultar a extração de informações através dos dados analisados, já que o impacto das informações faltantes nos resultados dos estudos nem sempre é conhecido (MCKNIGHT *et al.*, 2007).

Um dos fatores importantes para o aumento da ocorrência de valores faltantes é advindo de um procedimento relativamente comum, no qual eles substituem valores ruidosos - aqueles que apresentam um desvio significativo do valor real - na base de dados (MYRTVEIT; STENSRUD; OLSSON, 2001). Há vários tipos de dados ruidosos, dentre os principais: incorretos, duplicados e inconsistentes (WU; WUN; CHOU, 2004).

Como uma considerável parte das técnicas estatísticas foram projetadas para analisar dados completos, objetiva-se tratar os dados faltantes de forma a torná-los aptos de serem analisados por técnicas já consolidadas e tornando a inferência sobre os dados mais precisa (PEREIRA, 2014). Por outro lado, tendo conhecimento de algumas particularidades nos dados, tais como mecanismos geradores dos dados faltantes, padrão e quantidade, consegue-se encontrar

o melhor método dentre um conjunto existente de vários métodos para o tratamento de dados faltantes (VERONEZE, 2011).

Nesse sentido, um modelo bem treinado e validado não é o único fator relevante para uma boa previsão de valores futuros, mas esta também depende de um bom pré-processamento dos dados já que valores faltantes é um fator proibitivo na utilização de certas metodologias, além de afetar o desempenho de outras (SORJAMAA *et al.*, 2010).

2.2 Tipos de dados faltantes

Para que não ocorram inconsistências na análise dos dados, é essencial entendermos os motivos pelos quais surgiram estes dados faltantes, antes da aplicação de qualquer método. Os dados faltantes podem ser causados por combinações de três motivos: processos aleatórios, processos mensuráveis e processos não mensuráveis, sendo que, para esta última situação, os métodos de tratamento de dados faltantes geralmente não funcionam (GRAHAM; CUMSILLE; ELEK-FISK, 2003). Os dados faltantes podem ser de três tipos (RUBIN; LITTLE, 2019):

1) **Missing Completely at Random (MCAR):** para este mecanismo, que proporcionou a ocorrência dos dados faltantes foi um evento aleatório, também chamado de mecanismo *Missing Completely at Random* (VERONEZE, 2011), que diz respeito ao fato da omissão não estar relacionada às variáveis (dependentes ou independentes). Nessa situação, os valores faltantes para uma variável são uma simples amostra aleatória dos dados dessa variável ou, em outras palavras, a distribuição dos valores faltantes é de mesma natureza da dos valores observados (ZHANG, 2003).

   A causa de valores faltantes no banco de dados também pode ocorrer devido a uma variável não correlacionada com a variável que os possui. Nesse caso, o mecanismo também é MCAR, embora os dados faltantes existam devido a algum evento que possa não ser verdadeiramente aleatório (GRAHAM *et al.*, 1997). Um outro exemplo de ocorrência de um MCAR é quando o sensor de captura de dados para de funcionar por algum período de tempo ou quando se decide medir uma variável custosa apenas para um subconjunto aleatório da amostra.

   A principal vantagem de o mecanismo ser MCAR é que a causa que levou aos dados faltantes não precisa fazer parte da análise para controlar a influência destes nos resultados da pesquisa (GRAHAM *et al.*, 1997).

   Mesmo diante das vantagens associadas ao fato dos dados faltantes serem MCAR, vale ressaltar que a abordagem para se tratar deles deve ser cuidadosa, pois não é qualquer método dentre os capazes de lidar com dados faltantes que produzirá bons resultados.

2) **Missing at Random (MAR):** neste mecanismo, os dados faltantes são causados por alguma variável observada, disponível para análise e correlacionada com a variável que

possui dados faltantes (GRAHAM *et al.*, 1997). Sendo assim, podemos tratá-los como uma amostra aleatória simples dos dados para a variável que os contêm dentro de subgrupos definidos por valores observados da variável correlacionada, e a distribuição dos valores faltantes é a mesma que a distribuição dos valores observados dentro de cada subgrupo da variável correlacionada (ZHANG, 2003).

Uma situação de ocorrência do mecanismo MAR é o fato de pessoas com renda alta tenderem a não querer responder quantos televisores possuem em casa, por exemplo, indicando que o número de televisores está relacionado com a renda.

3) **Missing Not at Random (MNAR):** a omissão depende também do que não é observado. Nesse caso, as variáveis observadas não explicam completamente a omissão dos dados (PAES; POLETO, 2013).

Um exemplo clássico deste tipo de *missing* é também o valor da renda, em que pessoas com renda alta tendem a não querer informar a sua renda.

Os mecanismos MCAR e MAR também são chamados de ignoráveis, enquanto o mecanismo MNAR é denominado não-ignorável (GRAHAM *et al.*, 2009). A principal diferença entre eles, em relação a esta subdivisão, é que, pelo fato dos efeitos dos mecanismos ignoráveis nos modelos estatísticos estarem disponíveis para o analista de dados, então eles são considerados como mais fáceis de lidar (MCKNIGHT *et al.*, 2007). Dados faltantes do mecanismo MCAR geralmente não devem apresentar um impacto considerável na estimação dos parâmetros, já que os dados faltantes acontecem de maneira completamente aleatória. Por outro lado, quando o mecanismo é MAR, existe um processo sistemático subjacente à falta de dados que pode ser modelado por meio dos dados observados (MCKNIGHT *et al.*, 2007).

Em relação ao mecanismo não-ignorável, temos que o seu efeito é desconhecido e potencialmente perigoso, tendo em vista a não existência de nenhuma informação dentro do conjunto de dados que permita modelar e compreender a maneira com que os dados faltantes aconteceram. Portanto, ele deve ser modelado de forma a serem obtidas satisfatórias estimativas dos parâmetros de interesse.

Com base nestas informações, podemos dizer que o diagnóstico do mecanismo ajuda o pesquisador e o analista de dados a entender a natureza dos dados faltantes e o potencial impacto nos resultados dos estudos e nas interpretações destes (MCKNIGHT *et al.*, 2007). Mas como seria feito este diagnóstico? Para realizá-lo, primeiramente temos que verificar se o mecanismo é MCAR. Se não for, verifica-se se é MAR ou MNAR.

Para identificar se um mecanismo é MCAR, existe um único método formal, através do teste chi-quadrado (LITTLE, 1988). Por outro lado, se o mecanismo não for MCAR, é necessário saber se o mecanismo que criou os dados faltantes é relacionado com as informações conhecidas ou não. Infelizmente, não existe nenhum método formal para isso. Entretanto, há quatro situações em que podemos supor que o mecanismo é ignorável (SCHAFER; GRAHAM, 2002):

i) Quando algumas informações são colhidas de todos os elementos da base de dados, e outras informações adicionais são colhidas apenas de um subgrupo da amostra original, sendo que esse subgrupo é selecionado devido a alguma informação coletada para toda a amostra;

ii) Quando os pesquisadores podem substituir os itens da pesquisa que estejam com informações incompletas por outros cujos dados estão completos e que tenham as mesmas características;

iii) Em testes controlados aleatoriamente, em que o número de elementos (itens), nas diferentes intervenções, é desigual ou desbalanceado devido a causas inesperadas e não pelo motivo de haver um processo sistemático; e

iv) Quando temos informações a respeito de uma amostra e, posteriormente, coletamos informações adicionais referentes a um subgrupo selecionado de maneira aleatória ou com base nas informações colhidas previamente.

Dessa forma, se uma dessas situações acima ocorrer, o mecanismo pode ser considerado MAR e, caso contrário, deve ser considerado MNAR.

Neste trabalho, consideramos, inicialmente, a suposição de que os dados faltantes são MCAR para, em seguida, analisarmos como os métodos propostos se comportam com os dados MAR e MNAR.

2.3 Padrão e quantidade dos dados faltantes

Outro fator importante antes de se efetuar a escolha do procedimento mais adequado para tratar os dados faltantes é a análise da ocorrência de um determinado padrão entre eles, o qual descreveria quais valores foram observados e quais valores estão faltantes em uma matriz de dados (RUBIN; LITTLE, 2019). Dessa forma, conseguiríamos identificar se há ou não consistência no modo pelo qual os dados não foram observados (MCKNIGHT *et al.*, 2007).

Basicamente, dizemos que o padrão indica se os dados faltantes ocorrem de forma não-estruturada ou sistemática. Os dados faltantes são não-estruturados quando existem múltiplos padrões de dados faltantes entre os elementos do estudo, apontando que o mecanismo pode ser aleatório. Por outro lado, se os dados faltantes são sistemáticos, ou seja, estruturados ou capazes de expressar alguma tendência, pode significar que o mecanismo que os causou não é aleatório (MCKNIGHT *et al.*, 2007).

Por fim, o padrão pode ser uma ferramenta no auxílio da identificação de quais métodos são factíveis para o tratamento de um determinado conjunto de dados com informações faltantes.

Outro aspecto importante a ser considerado é a quantidade dos dados faltantes. Esse conceito pode atrelar-se à quantidade de elementos que possuem dados faltantes; de atributos

ou variáveis que possuem dados faltantes; de valores faltantes em um atributo específico; de valores faltantes em um conjunto específico de atributos e, enfim, de valores faltantes de todo o conjunto de dados. Sua importância está relacionada à eficácia dos processos de estimação que se pretende usar (VERONEZE, 2011).

Além disso, também podemos relacionar a quantidade de dados faltantes com a precisão das estimativas dos parâmetros, pois quanto maior essa quantidade, mais difícil será alcançar resultados satisfatórios de estimação.

Outro fator que merece atenção é a qualidade dos valores faltantes, pois uma grande quantidade de dados faltantes, sob um mecanismo ignorável, pode ser bem mais facilmente tratada do que uma pequena quantidade sob o mecanismo MNAR (MCKNIGHT *et al.*, 2007).

2.4 A seleção do método apropriado

Ao nos depararmos com um conjunto de dados com informações faltantes, há alguns procedimentos que podemos adotar para nos auxiliar no processo de seleção do método apropriado para lidarmos com eles.

O primeiro passo é identificar quais atributos ou variáveis do conjunto de dados são realmente relevantes à análise que será feita e, desta maneira, simplificar o diagnóstico e o tratamento dos dados faltantes, devido à redução do número de atributos ou variáveis que serão analisados.

O segundo passo consiste em especificar em que nível a análise será feita, caso os dados possuam uma estrutura hierárquica, como, por exemplo, micro e macro unidades, pois se a análise for feita no nível macro, dados faltantes no nível micro podem não ser tão relevantes (e vice-versa).

O terceiro passo é realizar o diagnóstico dos dados faltantes, isto é, identificar o mecanismo, o padrão e a sua quantidade, já que isto proporcionará a compreensão sobre os mesmos e, consequentemente, auxiliará no processo de tomada de decisão (VERONEZE, 2011).

Enfim, é imprescindível ter algum domínio acerca dos requisitos e suposições que as diversas técnicas que lidam com dados faltantes englobam, sendo que muitos desses requisitos e suposições estão notoriamente relacionados com as informações e características dos dados faltantes que foram levantadas anteriormente. O próximo capítulo irá formalizar alguns dos principais métodos de tratamento de dados faltantes.

CAPÍTULO

3

MÉTODOS PARA DADOS FALTANTES

No geral, nos casos em que os dados são MAR (*Missing at Random*) e MCAR (*Missing Completely at Random*), é seguro remover os dados com valores ausentes, dependendo de suas ocorrências, por se tratarem de mecanismos ignoráveis, enquanto que, no caso em que os dados são MNAR (*Missing not at Random*), caracterizado por ser um mecanismo não-ignorável, remover observações com valores ausentes pode produzir um viés no modelo. Portanto, temos que ter muito cuidado antes de remover as observações ou imputá-las, já que a imputação não fornece necessariamente melhores resultados, como veremos a seguir.

Neste capítulo, vamos apresentar os métodos mais utilizados para lidar com informações faltantes, subdividindo-os em **métodos de deleção**, **métodos de imputação única**, **máxima verossimilhança**, **métodos de imputação múltipla** e **métodos de aprendizado de máquina**.

3.1 Métodos de deleção

Os principais métodos de deleção são: *listwise*, *pairwise*, caso completo ponderado e descarte da variável com dados faltantes.

3.1.1 Listwise

Esta técnica consiste em eliminar todos os elementos com qualquer quantidade de dados faltantes nas variáveis. Em seguida, aplicam-se métodos convencionais de análise de conjuntos de dados completos, sendo, dessa forma, também conhecida como análise de casos completos (FICHMAN; CUMMINGS, 2003).

Há duas grandes vantagens para a aplicação da técnica de eliminação *listwise* (ALLISON, 2001):

a) ela pode ser usada para qualquer tipo de análise estatística, desde modelagem de equações

estruturais à análise de modelos log-lineares;

b) métodos computacionais especiais não são necessários.

Apenas os elementos que não tem dados faltantes sobre todas as variáveis dependentes e independentes são levados em consideração para análise (RIBEIRO, 2015), resultando em uma redução no tamanho da amostra original, mesmo quando há um pequeno número de variáveis em uma análise. Isso porque cada observação pode ter um valor faltante para apenas uma variável e não necessariamente para todas as variáveis (FICHMAN; CUMMINGS, 2003).

Se os dados são MCAR, especificamente, a deleção *listwise* pode assumir algumas propriedades atrativas em termos estatísticos, pois, neste mecanismo de dados faltantes, a amostra reduzida será uma sub-amostra aleatória da amostra original. Isto implica que, para qualquer parâmetro de interesse, se as estimativas forem não enviesadas para o conjunto de dados total, eles também serão não enviesados para o conjunto de dados obtido via exclusão por *listwise*. Vale ressaltar, igualmente, que os erros padrões e as estatísticas dos testes obtidos com o conjunto de dados excluídos via *listwise* serão tão apropriados como eles teriam sido no conjunto de dados completo. Por outro lado, como menos informação é utilizada no método *listwise*, o erro padrão geralmente será maior no conjunto de dados excluídos por esta técnica (ALLISON, 2001).

3.1.2 Pairwise

Esta técnica caracteriza-se por deletar apenas as observações com dados faltantes nas variáveis que serão necessárias para a análise, causando perda clara de informação que está disponível nos dados eliminados. Este método também é conhecido como análise de casos disponíveis (RIBEIRO, 2015) e é bastante parecido com o *listwise*, sendo a principal diferença o fato de que o *pairwise* não descarta os dados em nível de observação mas sim em nível de variáveis de interesse. Este é um procedimento alternativo para análises univariadas, pois ele inclui todas as observações em que a variável de interesse está presente (RUBIN; LITTLE, 2019).

A eliminação *pairwise* é uma alternativa simples que pode ser usada por muitos modelos lineares, dentre estes, regressão linear, análise fatorial e modelos mais complexos de equações estruturais, sendo seu principal objetivo calcular cada um destes resumos estatísticos utilizando todos os casos que estão disponíveis (ALLISON, 2001). Além disso, esta técnica é muitas vezes oferecida em pacotes de análise estatística que é aplicado para o cálculo da estatística descritiva (GRAHAM; HOFER; PICCININ, 1994).

Como este método descarta os elementos em nível de variável e não em nível de observação, uma de suas desvantagens é que a amostra-base muda de variável para variável, mediante o padrão dos dados faltantes. Logo, pode-se presumir que, para grandes bases de dados com diversos padrões de dados faltantes, os casos que provêem dados para uma variável podem ser

completamente diferentes dos casos que provêem dados para outra variável. Isso implica em vários problemas, dentre os principais:

i) Analisar a matriz de covariância ou a de correlação das variáveis, as quais são normalmente singulares ou indeterminadas (MCKNIGHT *et al.*, 2007);

ii) Calcular os erros padrão ou qualquer outra medida de incerteza (SCHAFER; GRAHAM, 2002), pois elas podem ficar subestimadas ou superestimadas.

Se os dados são MCAR, temos que o método *pairwise* possibilita estimativas consistentes dos parâmetros de interesse (ALLISON, 2001). Caso contrário, as estimativas podem ser seriamente enviesadas. Para o caso do mecanismo ser MCAR, amostras-bases diferentes são aceitáveis para estimativas de média e variância, mas não para estimativas de covariância e correlação (RUBIN; LITTLE, 2019).

Enfim, quando o mecaniso é MCAR e as correlações são modestas, temos que o método *pairwise* apresenta-se mais eficiente que o método *listwise* (KIM; CURRY, 1977). Porém, de um modo geral, nenhum desses dois métodos produz resultados satisfatórios (RUBIN; LITTLE, 2019).

3.1.3 Caso completo ponderado

O método caso completo ponderado é uma extensão do método *listwise* (caso completo) e se baseia nos dados observados para associar um valor (chamado peso) aos casos completos, com o objetivo de ajustar o viés (RUBIN; LITTLE, 2019). Ele busca amenizar os problemas oriundos do método *listwise*, tamanho menor da amostra e diminuição do poder estatístico na análise dos dados (MCKNIGHT *et al.*, 2007).

Este é um método identificado como método de ajuste, porque o objetivo dos pesos é aproximar a distribuição dos casos completos da distribuição da amostra completa da população (VERONEZE, 2011). Os pesos são empregados para corrigir a variabilidade da amostra de dados completos e os erros padrão associados as estimativas dos parâmetros (SCHAFER; GRAHAM, 2002).

Com o intuito de produzir pesos adequados, para cada variável com dados faltantes, é essencial que, a partir dos dados observados, sejam estimadas as probabilidades de cada possível resposta acontecer. Dessa forma, este método pode ser enfadonho quando existem muitas variáveis com dados faltantes e que possuem diferentes probabilidades de resposta e/ou quando existem muitos padrões de dados faltantes (MCKNIGHT *et al.*, 2007). Como este cenário é comum, este método é indicado somente para algumas condições: quando existem poucos padrões de dados faltantes e quando as probabilidades das respostas são conhecidas e relativamente uniformes entre as variáveis (MCKNIGHT *et al.*, 2007).

3.1.4 Descarte da variável com dados faltantes

Alguns pesquisadores preferem descartar as variáveis que possuem dados faltantes, durante o desenvolvimento de modelos preditivos. A principal vantagem deste tratamento é que desvinvula o modelo de futuras mudanças no perfil dos *missings*. Por exemplo, em uma situação em que um campo de preenchimento facultativo se torna obrigatório, o modelo se manteria intacto. Por outro lado, ao aplicarem tal método, pode-se correr o risco de não considerar uma informação valiosa que incrementaria o desempenho preditivo do modelo. Em uma situação extrema, é possível que todas as principais variáveis explicativas do modelo contenham dados faltantes, assim a sua não utilização inviabilizaria a construção do modelo (ASSUNÇÃO, 2012).

3.2 Métodos de imputação única

Imputação refere-se a um termo genérico para o preenchimento de dados faltantes com valores plausíveis (SCHAFER, 1997), podendo seus métodos serem de imputação única (IU) ou de imputação múltipla (detalhes das técnicas referentes à imputação múltipla serão vistos na próxima seção).

O objetivo principal dos métodos de IU é imputar um valor para cada dado faltante da base de dados e, então, analisá-la como se não houvesse dados faltantes (MCKNIGHT *et al.*, 2007). Ao contrário dos métodos de deleção apresentados nas seções anteriores, estes métodos visam predizer os valores faltantes. As estimativas dos parâmetros de interesse ficam em segundo plano, mas podem ser facilmente calculadas, visto que os métodos de IU produzem bases de dados completas, as quais podem ser analisadas através de procedimentos analíticos convencionais (VERONEZE, 2011).

3.2.1 Imputação por constantes

Dentre os métodos de IU, os métodos de imputação por constantes são os mais comuns (MCKNIGHT *et al.*, 2007). De forma geral, esses métodos substituem todos os valores faltantes de uma variável por um único valor (uma constante), por exemplo, imputação de zeros, imputação da média ou mediana da variável que possui os dados faltantes.

A técnica mais simples, dentre os exemplos citados, é o método de imputação de zeros, que consiste, basicamente, em trocar os valores faltantes por zero, desde que o zero seja um valor plausível para o conjunto de dados. Dessa forma, se o motivo dos dados faltantes está ligado com maus resultados e os maus resultados estão associados a valores próximos de zero, a imputação de zeros pode não ter qualquer influência nos resultados da análise (MCKNIGHT *et al.*, 2007).

Por outro lado, como já foi dito no capítulo anterior, existem diversas razões que levam a ocorrência de dados faltantes, sendo que a considerável parte delas não justifica que a pior (ou mais conservadora) resposta seja considerada. Portanto, de modo geral, esta técnica não produz

bons resultados (MCKNIGHT *et al.*, 2007), além do fato de que, quanto maior a quantidade de dados faltantes, mais os resultados tendem a serem piores, devido à maior quantidade de zeros imputados.

Quanto à imputação pela média, temos que é um método comum (MYRTVEIT; STENSRUD; OLSSON, 2001) e bastante utilizado (BROWN; KROS, 2003) pela sua facilidade de implementação. Nesta técnica, a média dos valores de uma variável que contém dados faltantes é usada para preencher os seus dados faltantes. No caso de variáveis categóricas, a moda é usada no lugar da média (FARHANGFAR; KURGAN; PEDRYCZ, 2004). Apesar da fácil implementação, na maioria dos casos este método não se apresenta eficaz para o tratamento das informações faltantes, pois, com sua aplicação, os valores extremos ficam sub-representados, implicando em perda de variabilidade, ou seja, a variância das variáveis com dados faltantes é subestimada. Dessa forma, o efeito prejudicial deste método é reduzido somente em bases de dados com uma pequena porcentagem de dados faltantes (MCKNIGHT *et al.*, 2007). É importante ressaltar, igualmente, que a média é a melhor medida de tendência central para variáveis normalmente distribuídas, sendo assim, quando a distribuição normal não se verifica, os resultados deste método podem ser bastante insatisfatórios.

Por fim, a imputação pela mediana é bem parecida com a imputação pela média, apresentando as mesmas vantagens e desvantagens. Porém, a imputação da mediana se apresenta como uma alternativa melhor para variáveis que não são normalmente distribuídas, pois a mediana representa melhor a tendência central de uma distribuição que possui muitos valores atípicos.

3.2.2 Hot-deck e cold-deck

As técnicas baseadas em *Hot-Deck* (HD) preenchem um valor faltante de uma variável referente a um elemento da pesquisa a partir do(s) valor(es) observado(s) para esta mesma variável em outro(s) elemento(s) do mesmo conjunto de dados (SCHAFER; GRAHAM, 2002).

Existem vários métodos HD e o mais simples consiste em imputar dados a partir de um elemento completo escolhido aleatoriamente. Outro método de aplicação bastante básica se baseia em escolher um elemento semelhante ao elemento com dado faltante com base em alguma informação e copiar o valor faltante desse elemento parecido. Para ilustrar, se uma informação faltante está num elemento que se refere a uma pessoa do sexo feminino, então algum elemento será escolhido aleatoriamente dentre os elementos que se referem às pessoas do sexo feminino para fornecer o valor a ser imputado (MCKNIGHT *et al.*, 2007).

Desenvolvendo-se o método de HD mais elaboradamente, temos que, para todo elemento que contém dados faltantes, o(s) elemento(s) mais semelhante(s) é(são) encontrado(s) e, então, os dados faltantes são imputados a partir desse(s) elemento(s) (LAKSHMINARAYAN; HARP; SAMAD, 1999; FARHANGFAR; KURGAN; PEDRYCZ, 2004; BROWN; KROS, 2003). Se nos depararmos com a situação em que um elemento dito similar também contém informações

faltantes para as mesmas variáveis consideradas, então ele é descartado, e outro elemento similar é procurado para a realização do método (FARHANGFAR; KURGAN; PEDRYCZ, 2004; FARHANGFAR; KURGAN; PEDRYCZ, 2007).

As principais vantagens de se optar pelos métodos HD são: (i) simplicidade conceitual; (ii) bom nível de medição de variáveis; e (iii) como em todos os métodos de imputação, uma base de dados completa é gerada, e esta, então, pode ser analisada através de procedimentos analíticos convencionais. Por outro lado, uma das desvantagens de HD é a dificuldade em definir o que é similar, pois, embora existam várias métricas que podem ser utilizadas, como a distância euclidiana e a correlação de Pearson, a escolha da métrica certamente vai influenciar no desempenho de HD (BROWN; KROS, 2003).

Outra desvantagem é a subestimação dos erros padrão (que pode implicar em erros do Tipo I, significando que se pode aceitar uma afirmação sem evidências suficientes), devido à diminuição da variabilidade nas variáveis com dados faltantes, já que tendem a serem imputados valores distantes dos extremos da distribuição (MCKNIGHT *et al.*, 2007).

Quanto aos métodos *Cold-Deck* (CD), temos que eles são muito similares aos métodos HD. O que os difere é que no HD os dados utilizados para a substituição dos dados faltantes estão no próprio conjunto de dados, enquanto que no CD, estão em outro conjunto de dados (LAKSHMINARAYAN; HARP; SAMAD, 1999; ACUNA; RODRIGUEZ, 2004). Como exemplo, podemos citar o fato de que, numa pesquisa longitudinal, essa fonte de dados externa pode ser a medição anterior ou a posterior. Vale ressaltar a importância de se certificar de que a fonte de dados externa contenha valores factíveis para o preenchimento dos dados faltantes (BROWN; KROS, 2003) e assim, através dessa fonte, o método CD procura diminuir o problema da perda de variabilidade, visando, com isso, evitar os erros do Tipo I (MCKNIGHT *et al.*, 2007).

3.2.3 Outras técnicas de imputação única

Exitem diversas outras técnicas de IU além das já citadas. Como exemplo, podemos mencionar: Imputação de médias condicionadas (CMI – do inglês *Conditional Mean Imputation*), Próximo valor carregado para trás (NVCB – do inglês *Next Value Carried Backward*) e Último valor carregado para frente (LVCF – do inglês *Last Value Carried Forward*) (MCKNIGHT *et al.*, 2007).

Quanto ao método CMI, temos que, ao contrário do método de imputação de médias, em que, para calcular a média de uma variável utiliza-se todos os registros observados desta variável, ele calcula a média para diferentes subgrupos formados a partir de variáveis de classificação. Por exemplo, se sexo é a única variável de classificação e tem-se um dado faltante para uma pessoa do sexo feminino, este será substituído pela média calculada dentro do grupo de pessoas do sexo feminino. Logicamente, quanto mais fraca a relação entre as variáveis de classificação e a variável com dados faltantes, mais este método se aproxima de um método de imputação de

valores aleatórios (MCKNIGHT *et al.*, 2007).

Por fim, em relação aos métodos LVCF ou NVCB, eles são possíveis de serem utilizados em estudos longitudinais, em que o método LVCF consiste, basicamente, em substituir o valor faltante de uma variável para um elemento pelo valor dessa mesma variável na medição anterior desse elemento, enquanto que o método NVCB se caracteriza por substituir o valor faltante de uma variável para um elemento pelo valor dessa variável na medição posterior desse mesmo elemento (VERONEZE, 2011).

3.3 Máxima verossimilhança sem estrutura de regressão entre as variáveis

Os métodos de Máxima Verossimilhança (MV) são elaborados de acordo com um modelo e as estimativas dos parâmetros são calculadas a partir dos dados observados, das relações existentes entre as variáveis com valores observados e também das restrições impostas pela suposição do modelo de distribuição (MCKNIGHT *et al.*, 2007).

Nesta técnica, o objetivo principal não é preencher os valores faltantes, mas sim estimar os parâmetros de interesse, diferindo-se, assim, dos métodos de imputação. Temos que, os métodos baseados em modelo tratam os dados como se todos os valores tivessem sido observados, pois eles combinam os dados com um modelo teórico (por exemplo, a distribuição normal multivariada) e, assim, são estimados os parâmetros de interesse (MCKNIGHT *et al.*, 2007).

O princípio básico dos métodos MV é escolher como estimativa dos parâmetros aqueles valores que, se verdadeiros, maximizariam a probabilidade de observar o que, de fato, foi observado (ALLISON, 2001).

Quanto à realização de um procedimento MV, temos que é necessário se considerar que os dados são gerados por um modelo descrito pela função de densidade $f(A|\theta)$, em que A são os dados e θ é um conjunto de parâmetros desconhecidos que rege a distribuição de A, do qual sabe-se apenas estar situado no espaço paramétrico Ω_θ (RUBIN; LITTLE, 2019). Salvo indicação contrária, consideram-se intervalos adequados para os elementos de θ, por exemplo: o espaço dos reais para médias, os reais positivos para variâncias e o intervalo $[0,1]$ para probabilidades (RUBIN; LITTLE, 2019). Logo, dado o modelo considerado e uma vez estimado o vetor de parâmetros θ, $f(A|\theta)$ pode ser utilizada para amostrar valores faltantes (RUBIN; LITTLE, 2019), (ALLISON, 2001).

Suponha que exista uma variável Y no conjunto de dados A e que se objetiva estimar o parâmetro θ. Se $f(y|\theta)$ é a probabilidade (ou densidade de probabilidade) de observar um único valor de Y dado algum valor de θ, a verossimilhança para uma amostra de n observações

independentes é definida por (ALLISON, 2001):

$$L(\theta|y) = \prod_{i=1}^{n} f(y_i|\theta). \tag{3.1}$$

Em alguns problemas é mais atrativo, até mesmo computacionalmente, trabalhar com a função $l(\theta|A)$ (log-verossimilhança), que é o logaritmo natural (log) da função de verossimilhança, pois, como o logaritmo é uma função monotonicamente crescente, o logaritmo da função alcança o seu valor máximo nos mesmos pontos que a função original. Portanto, a função $l(\theta|A)$ pode ser usada no lugar de $L(\theta|A)$ na estimação da máxima verossimilhança. Vale lembrar, igualmente, que encontrar o máximo de uma função sempre envolve o cálculo de derivadas e isso se torna mais fácil quando a função que está sendo maximizada é $l(\theta|A)$ (RUBIN; LITTLE, 2019).

Dessa forma, concluímos que o objetivo dos métodos de MV é encontrar o valor de $\theta \in \Omega_\theta$ que maximiza a função de verossimilhança $L(\theta|A)$, ou, equivalentemente, $l(\theta|A)$. Em alguns casos, é possível encontrar mais de uma estimativa MV, porém, para muitos modelos importantes, a estimativa MV é única (RUBIN; LITTLE, 2019).

Quando temos um mecanismo de dados faltantes ignorável, conseguimos obter a verossimilhança simplesmente pela soma, para variáveis categóricas ou discretas, ou pela integral, para variáveis contínuas, da verossimilhança usual sobre todos os possíveis valores dos dados faltantes (ALLISON, 2001). Suponha, por exemplo, que se tentou coletar dados sobre duas variáveis, X e Y, para uma amostra independente de n observações. Para as primeiras m observações, ambas as variáveis foram observadas, entretanto, para as $n-m$ observações restantes, somente Y foi observada. Para uma única observação com dados completos, a função de densidade é descrita por $f(x,y|\theta)$, em que θ é um conjunto de parâmetros desconhecidos que governam a distribuição de X e Y. Partindo-se da suposição de que X é discreta, a função de densidade para um caso com dados faltantes em X é a distribuição marginal de Y:

$$g(y|\theta) = \sum_{x} f(x,y|\theta). \tag{3.2}$$

Como consequência, a verossimilhança para toda a amostra é:

$$L(\theta|\mathbf{x},\mathbf{y}) = \prod_{i=1}^{m} f(x_i,y_i|\theta) \prod_{i=m+1}^{n} g(y_i|\theta). \tag{3.3}$$

Conhecidas pela sua eficiência, as técnicas MV estão implementadas em vários softwares estatísticos (MYRTVEIT; STENSRUD; OLSSON, 2001), já que elas possuem várias características desejáveis, tais como: i) produzir estimativas aproximadamente não-enviesadas para grandes amostras; ii) os erros padrão são, ao menos, tão pequenos quanto os erros padrão produzidos por outro método consistente e iii) em amostragens repetidas, as estimativas têm, aproximadamente,

uma distribuição normal, o que pode ser utilizado para calcular intervalos de confiança. Além disso, lembremos que este método é considerado eficiente sob o mecanismo ignorável e também que, quanto maior a amostra, maior a tendência de se obter resultados satisfatórios (MCKNIGHT *et al.*, 2007), (ALLISON, 2001).

Enfim, pelo fato dos métodos MV exigirem que seja considerado um modelo, é de inestimável importância que esta escolha seja bem feita, pois, caso contrário, estes métodos podem ser acometidos por uma diminuição drástica de precisão. Logo, esta é a principal desvantagem referente a esta técnica de se lidar com dados faltantes, já que nem sempre o analista de dados vai conseguir propor um modelo adequado (ALLISON, 2001).

3.3.1 O algoritmo EM

De acordo com Little & Rubin (2002), o algoritmo EM (do inglês, *Expectation-Maximization*) é definido como um método geral para obter estimativas MV em bases de dados incompletas, pois, devido ao fato das estimativas MV serem difíceis de ser calculadas para bases de dados complexas, torna-se necessário um procedimento para reduzir esta dificuldade, e é este o intuito do algoritmo EM (MCKNIGHT *et al.*, 2007; RUBIN; LITTLE, 2019).

O nome EM vem de seus dois principais passos de realização: Esperança e Maximização. Como estes dois passos se alternam até a convergência, temos que este, então, é um algoritmo iterativo. Podemos explicar, de forma sucinta, o algoritmo da seguinte maneira:

i) Escolher valores iniciais para os parâmetros do modelo considerado (por exemplo, médias e matriz de covariância para o modelo normal multivariado), em que estes valores iniciais podem ser obtidos através das fórmulas convencionais, usando *Listwise Deletion* ou *Pairwise Deletion* (ALLISON, 2001).

ii) Faça até a convergência:

   - **Esperança:** imputar valores para os dados faltantes baseando-se nos valores dos parâmetros. Por exemplo: considere que em uma análise de dados trabalha-se com a distribuição normal multivariada, na qual a base de dados possui 4 variáveis (A_1, A_2, A_3, A_4) e que exista alguns dados faltantes em A_3 e A_4 (sendo que o padrão de dados faltantes é arbitrário). Os valores podem ser imputados por meio da distribuição condicional conjunta de A_3 e A_4 dado A_1 e A_2 (ALLISON, 2001);
   - **Maximização:** estimar novos valores dos parâmetros. Retornando ao exemplo do item anterior, a média pode ser calculada utilizando a fórmula convencional, mas a variância e a covariância devem levar em consideração um termo adicional que corresponde aos resíduos das mesmas. A adição do termo residual corrige a subestimação da variância que ocorre nos esquemas mais convencionais de imputação (ALLISON, 2001).

Sabemos que o método convergiu quando a diferença entre os valores estimados dos parâmetros ou quando a diferença no valor da log-verossimilhança entre duas iterações consecutivas é menor que um limiar pré-estabelecido.

Como o algoritmo EM é um método baseado em MV, ele possui propriedades desejáveis quando o mecanismo dos dados faltantes for ignorável, produzindo estimativas MV muito satisfatórias em termos de consistência e eficiência para grandes amostras. No entanto, o procedimento EM tende a subestimar os erros padrão utilizados em testes de hipóteses, podendo resultar em erros do Tipo I (MCKNIGHT *et al.*, 2007; GRAHAM *et al.*, 2009).

Outras duas desvantagens sobre o algoritmo EM são: (i) em alguns casos, com grandes frações de informações faltantes, ele pode ter um processo de convergência muito lento; e (ii) em algumas aplicações, o passo de maximização não tem uma formulação computacional simples (RUBIN; LITTLE, 2019). Existem algumas extensões do algoritmo EM que tentam amenizar esses inconvenientes, como por exemplo, o algoritmo ECM (MENG; RUBIN, 1993).

Outras desvantagens do algoritmo EM é que, nos casos em que a função de verossimilhança (ou log-verossimilhança) for multimodal, ele não garante a convergência para o máximo global e pode ficar preso em máximos locais. Esse algoritmo também é sensível ao valores iniciais dos parâmetros, sendo importante uma escolha adequada desses valores para alcançar o máximo global.

3.4 Imputação múltipla

Os métodos de imputação múltipla (IM) surgiram como uma alternativa flexível aos métodos MV para uma grande variedade de problemas de dados faltantes (RUBIN, 1987; SCHAFER; GRAHAM, 2002). Esse método se caracteriza por três características: i) oferece estimativas de parâmetros confiáveis (incluindo erros padrão); ii) permite a predição da informação faltante e o seu impacto nas estimativas dos parâmetros e iii) pode ser aplicado nas mais diversas situações envolvendo dados faltantes (MCKNIGHT *et al.*, 2007).

A principal diferença entre os métodos IM e os métodos IU é que, enquanto os métodos de imputação única substituem cada valor faltante por um único valor, as técnicas de imputação múltipla substituem por x valores, com $x \geq 2$. Dessa forma, são formadas x bases de dados completas, que, portanto, podem ser analisadas através de procedimentos convencionais. Em seguida, os resultados dessas análises são agregados e, por fim, a informação faltante pode ser computada.

Como IM é o nome dado a uma família de métodos, é necessário saber selecionar o método adequado, sendo que, o processo de seleção do método IM adequado deve ser guiado pela suposição sobre a distribuição dos dados, como também pelos tipos de dados (contínuos, discretos e/ou categóricos) presentes na base de dados (LAKSHMINARAYAN; HARP; SAMAD,

1999). Isso porque em procedimentos IM, um valor faltante é substituído com base nos valores observados e em um erro que é adicionado para garantir a conformidade com a distribuição considerada (MCKNIGHT *et al.*, 2007).

Em geral, a distribuição mais comumente utilizada para as variáveis do conjunto é a normal (ALLISON, 2001), pois, mesmo quando algumas variáveis não são normais, esse modelo se mostra apto de ser aplicado (SCHAFER, 1997). Além disso, muitas vezes é possível transformar os dados de uma variável para que eles se ajustem em uma distribuição normal, quando os mesmos não seguem esta distribuição.

Enfim, este método se realiza por meio dos seguintes passos: imputação, análise das bases de dados geradas e agregação dos resultados, explicados com mais detalhes nas subsseções a seguir.

3.4.1 Imputação

Consiste no passo fundamental da técnica de IM (ZHANG, 2003) e é semelhante à IU, porém são realizadas $x \geq 2$ imputações para cada valor faltante. Não existem restrições quanto ao método escolhido para realizar as x imputações, os quais podem ser, por exemplo, estimativas MV, *Hot-Deck* ou *Cold-Deck*. Por outro lado, utilizar mais de um método não é, geralmente, sensato, porque eles produziriam resultados diferentes, influenciando na estimativa da informação faltante, pois produziriam resultados piores com o aumento da variabilidade. A maioria dos especialistas em IM recomenda um procedimento iterativo que não é limitado somente a um grupo específico de valores imputados (como a média, por exemplo). Ao invés disso, um processo aleatório é preferível, de modo que os valores sejam únicos em cada conjunto imputado, mas compartilhem uma relação comum subjacente aos dados (MCKNIGHT *et al.*, 2007). Essa variação aleatória é introduzida para evitar que as variâncias das variáveis sejam subestimadas e prejudiquem as inferências estatísticas (ALLISON, 2001; MCKNIGHT *et al.*, 2007).

Além disso, enquanto a variação aleatória introduzida nas imputações pode eliminar o viés que é endêmico às imputações determinísticas, as imputações múltiplas podem superestimar os erros padrão e reduzir a probabilidade de erros do Tipo I (ALLISON, 2001; MCKNIGHT *et al.*, 2007; SCHAFER; GRAHAM, 2002). Isso acontece devido ao componente aleatório introduzido na imputação, que possibilita que as estimativas dos parâmetros de interesse sejam levemente diferentes em cada base de dados imputada.

Enfim, é igualmente importante ressaltar que qualquer técnica utilizada para tratar os dados faltantes é notoriamente influenciada pelo tipo de dados envolvido (LAKSHMINARAYAN; HARP; SAMAD, 1999). Sendo assim, a escolha do método de imputação deve levar em consideração os tipos de dados presentes na base.

3.4.2 Análise e agregação dos resultados

Para estimar os modelos estatísticos de interesse para o estudo, as $x \geq 2$ bases de dados geradas são usadas em análises convencionais. Cada uma das x bases de dados completas é analisada individualmente e, logicamente, como são realizadas x análises, serão geradas x estimativas de cada um dos parâmetros de interesse. Vale ressaltar que não existe nenhuma restrição quanto ao tipo de análise estatística, de modo que análises univariadas ou multivariadas são igualmente aplicáveis (MCKNIGHT *et al.*, 2007).

Exemplificando, vamos supor que exista uma base de dados com quatro variáveis: preço, peso, cor e claridade de um diamante, sendo que o preço do diamante é a variável dependente. Agora, vamos adotar a regressão múltipla como o modelo estatístico de interesse para o estudo em questão. Logo, após realizar a imputação, é possível analisar cada uma das bases de dados através da regressão múltipla e, assim, para cada base de dados completa, encontra-se o coeficiente de regressão de cada uma das variáveis independentes e também o intercepto, além dos erros padrão. Nesse exemplo, os coeficientes de regressão e o intercepto são os parâmetros de interesse.

O penúltimo passo de realização do método de imputação múltipla é agregar os resultados das análises feitas nas $x \geq 2$ bases de dados, gerando, assim, estimativas globais para os parâmetros de interesse e para os erros padrão.

Uma maneira simples para se gerar uma estimativa global para um parâmetro de interesse Q é através da média das estimativas produzidas para as x bases de dados (MCKNIGHT *et al.*, 2007).

3.4.3 Algumas considerações sobre o IM

IM tem as mesmas vantagens de IU, entretanto, esta técnica está isenta de algumas desvantagens que acometem os métodos de IU, dentre elas: i) quando as $x \geq 2$ imputações são realizadas utilizando o mesmo modelo, os resultados das análises das $x \geq 2$ bases de dados podem ser facilmente combinados para criar uma inferência que reflete adequadamente a variabilidade da amostra e ii) quando as IM's são feitas a partir de mais de um modelo, a incerteza sobre o modelo correto é mostrada pela variação nas inferências válidas entre os modelos (RUBIN; LITTLE, 2019).

Agora, se comparados aos métodos MV, os métodos IM também estão isentos de algumas desvantagens, pois, ao contrário de MV, IM pode ser utilizado com praticamente qualquer tipo de dados e qualquer tipo de modelo, e a análise pode ser feita com softwares convencionais, sem nenhuma modificação. Além disso, vale ressaltar que IM é, provavelmente, menos sensível do que MV à escolha do modelo, já que o modelo é usado somente para a imputação dos valores e não para estimar os parâmetros (ALLISON, 2001).

Um dos principais inconvenientes da aplicação dos métodos de imputação múltipla, é que esta é uma técnica de implementação difícil. Porém, com o avanço da computação e com

a proliferação dos softwares para realizar IM, ela se tornou, rapidamente, o método padrão para manipular dados faltantes (ALLISON, 2001; MCKNIGHT *et al.*, 2007). Um exemplo de software que trabalha com IM é o *NORM*, um programa gratuito para Windows, que cria IM's para bases de dados incompletas, com padrão arbitrário, sob um modelo normal não estruturado (SCHAFER; GRAHAM, 2002).

Enfim, mesmo diante dos vários resultados satisfatórios que a IM oferece com relação aos métodos de deleção, imputação única e baseados em modelo, ela nem sempre é a melhor opção. Por exemplo, como a IM é um método de simulação, ela pode ser menos eficiente sob condições em que métodos baseados em modelo podem calcular os parâmetros de interesse diretamente do conjunto de dados incompleto (MCKNIGHT *et al.*, 2007).

3.5 Métodos de aprendizado de máquina

Visando tratar os casos de omissão, alguns métodos de aprendizado de máquina também têm sido propostos, destacando-se, dentre eles, o *Autoclass* e o *C4.5*.

Autoclass caracteriza-se como uma técnica de agrupamento usado para revelar a estrutura intrínseca nos dados. Podemos mencionar, como uma característica interessante do *Autoclass*, que ele procura por classes automaticamente, e tem limites que impedem dados de *over-fitting* (que é a memorização dos padrões, que tem como consequência um erro quadrático baixo na fase de treinamento, porém um erro quadrático alto na fase de teste) (LAKSHMINARAYAN; HARP; SAMAD, 1999). Por outro lado, o *C4.5* é um algoritmo de árvore de decisão para classificação e baseia-se na teoria de classificação Bayesiana, que poderia ser utilizada para prever diferentes atributos após uma simples sessão de aprendizagem. Isso permite que seu uso seja econômico em termos de tempo.

Outras técnicas também têm sido abordadas, que são as Redes Neurais MLP, *Random Forest Imputation*,*Weighted Imputacion with K-Nearest Neighbor - WKNNI*, *K-means Clustering Imputacion - KMI*, *Support Vector Machines Imputacion - SVMI*, *Singular Value Decomposition Imputacion - SVDI*, *K2*, *Data Augmentation - DA*, *BN - K2I*$_{\chi^2}$, *IBN - K2I*$_{\chi^2}$, algoritmo de biclusterização *SwarmBcluster* (JR; EBECKEN, 2002; VERONEZE, 2011; LUENGO; GARCÍA; HERRERA, 2012).

3.6 Modelo probabilístico para variáveis dicotômicas

Para o entendimento da construção de um modelo probabilístico para imputação de dados faltantes em duas variáveis dicotômicas, utiliza-se um caso particular em que se deseja classificar as respostas das unidades experimentais de acordo com duas variáveis binárias, Y_1 e Y_2, que podem assumir os valores 1 e 2. Essas variáveis podem representar, por exemplo, a categorização de cada unidade amostral com relação a duas questões de interesse ou a mesma questão medida

em duas ocasiões, originando um dos seguintes cenários: (i) classificação completa (em Y_1 e Y_2); (ii) classificação em Y_1 e omissão em Y_2; (iii) omissão em Y_1 e classificação em Y_2 e (iv) omissão completa (em Y_1 e Y_2) (POLETO, 2006).

Entretanto, apesar de o interesse inicial se concentrar apenas nas variáveis Y_1 e Y_2, a ocorrência de unidades amostrais nos cenários de omissão (ii), (iii) e (iv) sugere a utilização de uma terceira variável, W, com possíveis valores $1,2,3$ e 4, representativa dos diferentes padrões de omissão.

Considere que a distribuição do vetor aleatório (W,Y_1,Y_2) possui 16 parâmetros denotados por $\gamma_{tij} = P(W = t, Y_1 = i, Y_2 = j),\ i,j = 1,2,\ t = 1,2,3,4$, sendo que apenas 15 são linearmente independentes, pois $\sum_{t=1}^{4}\sum_{i=1}^{2}\sum_{j=1}^{2}\gamma_{tij} = 1$. Além disso, as probabilidades conjuntas γ_{tij} podem ser escritas como o produto das probabilidades marginais de (Y_1,Y_2) denotadas por θ_{ij}, pelas probabilidades condicionais de $W|Y_1,Y_2$, representadas por $\lambda_{t(ij)}$, ou seja,

$$\begin{aligned}\gamma_{tij} &= P(W = t, Y_1 = i, Y_2 = j) \\ &= P(Y_1 = i, Y_2 = j)P(W = t|Y_1 = i, Y_2 = j) = \theta_{ij}\lambda_{t(ij)}.\end{aligned} \tag{3.4}$$

Observemos que, no cenário $W = 1$, as variáveis Y_1 e Y_2 são observadas e, portanto, as probabilidades $\{\lambda_{1(ij)}\}$ devem ser interpretadas como probabilidades de ausência de omissão. Note também que, devido às restrições $\sum_{t=1}^{4}\lambda_{t(ij)} = 1, i,j = 1,2$, pode-se obter de maneira única as probabilidades condicionais de omissão, estabelecendo uma estrutura apenas para $\lambda_{t(ij)}, t = 2,3,4, i,j = 1,2$ e tomando $\lambda_{1(ij)} = 1 - \lambda_{2(ij)} - \lambda_{3(ij)} - \lambda_{4(ij)}, i,j = 1,2$. Isso demonstra que as probabilidades condicionais de ausência de omissão são funções das outras probabilidades condicionais de omissão.

Essa fatoração facilita a explicitação de modelos estruturais para as **probabilidades marginais de categorização**, de interesse primordial, e para as **probabilidades condicionais de omissão**, de interesse secundário. Pelo fato das probabilidades condicionais de omissão $\{\lambda_{t(ij)}\}$ estarem associadas a uma auto-seleção das unidades amostrais em algum dos padrões de omissão W, dado $Y_1 = i, Y_2 = j$, temos que os modelos estruturais propostos por meio da fatoração (3.4) são chamados de **modelos de seleção**.

Mediante o que foi apresentado no Capítulo 2 em relação às características referentes aos dados faltantes, dentre as quais podemos citar os mecanismo geradores dos mesmos e, de acordo com os métodos mais comumente empregados para lidar com eles citados neste capítulo e suas principais desvantagens, o objetivo específico desta pesquisa é desenvolver quatro metodologias que não dependam da remoção ou imputação de dados (divididas entre método com resolução analítica e métodos sem resolução analítica) e que obtenham performances inferenciais e preditivas satisfatórias mesmo sob o mecanismo MNAR de geração de dados

faltantes, considerado o mecanismo mais desafiador para os métodos já existentes, e para quaisquer distribuições que tenham os dados.

CAPÍTULO

4

MÉTODO BASEADO EM MODELO COM RESOLUÇÃO ANALÍTICA

Nesta capítulo, desenvolvemos a primeira proposta desta tese, que é o método de estimação baseado em modelos na presença de valores faltantes com solução analítica (MMORA). Em seguida, trabalhamos com um exemplo com resolução analítica baseado em um modelo de regressão linear múltipla.

4.1 O MMORA

Sejam $Y, X_1, ..., X_k$ variáveis aleatórias com função densidade de probabilidade conjunta $f(y, x_1, ..., x_k|\theta)$. Então, a função densidade de probabilidade de Y, dado $(X_1 = x_1, X_2 = x_2, ..., X_k = x_k)$, é definida para todos os valores $(x_1, ..., x_k)$ tal que $f(x_1, ..., x_k|\theta) > 0$ por:

$$f(y|x_1, ..., x_k, \theta) = \frac{f(y, x_1, ..., x_k|\theta)}{f(x_1, ..., x_k|\theta)}. \tag{4.1}$$

Considerando, agora, a distribuição condicional de $X_1|X_2, ..., X_k$, definida para todos os valores $x_2, ..., x_k$ tal que $f(x_2, ..., x_k) > 0$, temos que:

$$f(x_1|x_2, ..., x_k, \theta) = \frac{f(x_1, ..., x_k|\theta)}{f(x_2, ..., x_k|\theta)}. \tag{4.2}$$

Logo, da Equação (4.1) e (4.2), segue que:

$$\begin{aligned} f(y, x_1, ..., x_k|\theta) &= f(y|x_1, ..., x_k, \theta) f(x_1, ..., x_k|\theta) \\ &= f(y|x_1, ..., x_k, \theta) f(x_1|x_2, ..., x_k, \theta) f(x_2, ..., x_k|\theta). \end{aligned} \tag{4.3}$$

Por outro lado, a distribuição condicional de $Y, X_1|X_2,...,X_k$ nos diz que:

$$\frac{f(y,x_1,...,x_k|\theta)}{f(x_2,...,x_k|\theta)} = f(y,x_1|x_2,...,x_k,\theta). \tag{4.4}$$

Da Equação (4.3) e (4.4), segue que:

$$f(y,x_1|x_2,...,x_k,\theta) = f(y|x_1,...,x_k,\theta)f(x_1|x_2,...,x_k,\theta). \tag{4.5}$$

Seja Y a variável resposta (variável de interesse do experimento, sendo medida ou observada), e $X_1,...,X_k$ as variáveis explicativas (outras variáveis no experimento que afetam a resposta). Se a variável X_1 possuir valores faltantes, temos interesse na função densidade $f(y|x_2,...,x_k,\theta)$ para estes e, para obtê-la, basta integrarmos a Equação (4.5) em relação a x_1.

Assim, para dados completos, consideramos $f(y|x_1,...,x_k,\theta)$ e, para dados faltantes, $f(y|x_2,...,x_k,\theta)$, que depende da definição de $f(x_1|x_2,...,x_k,\theta)$. A principal ideia é, através de probabilidades condicionais, escrevermos um modelo que utilize as informações disponíveis, sem depender da imputação de dados. Observe que, diferente de um modelo de regressão linear tradicional, cujas variáveis explicativas são vistas como características fixas, as variáveis explicativas com valores faltantes são analisadas pelos método proposto como variáveis aleatórias para as quais definimos uma distribuição de probabilidade de acordo com suas naturezas. A esperança das variáveis explicativas com valores faltantes são então definidas como função das outras variáveis explicativas disponíveis.

Consideremos, agora, uma amostra de tamanho n de cada uma das variáveis $Y, X_1, ..., X_k$, $(y_1,...,y_n,x_{11},...,x_{1n},x_{21},...,x_{2n},...,x_{k1},...,x_{kn})$, em que algumas observações de X_1 podem não estar disponíveis (dados faltantes). Para a i-ésima observação, temos dois cenários diferentes em relação à densidade condicional:

a) Quando X_1 é observado, temos:

$$f(y_i|x_{1i},...,x_{ki},\theta); \tag{4.6}$$

b) Quando X_1 está faltante ou é desconhecido, temos:

$$\begin{aligned} f(y_i|x_{2i},...,x_{ki},\theta) &= \int f(y_i,x_{1i}|x_{2i},...,x_{ki},\theta)dx_{1i} \\ &= \int f(y_i|x_{1i},...,x_{ki},\theta)f(x_{1i}|x_{2i},...,x_{ki},\theta)dx_{1i} \\ &= E_{X_1|X_2,...,X_k}[f(y_i|X_{1i},...,x_{ki},\theta)], \end{aligned} \tag{4.7}$$

que é a definição da esperança condicional de $f(y_i|X_{1i},...,x_{ki},\theta)$ em relação à $X_1|X_2,...,X_k$.

Agora, seja $\delta_j(i)$, para j referente ao índice da variável que possui valores faltantes (neste caso $j = 1$), a função indicadora se a observação i é observada em relação à variável X_j, ou seja:

$$\delta_j(i) = \begin{cases} 1, & \text{se a observação } i \text{ é observada em relação a } X_j; \\ 0, & \text{se a observação } i \text{ é faltante em relação a } X_j. \end{cases} \tag{4.8}$$

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

$$\begin{aligned} L(\theta|\mathbf{y},\mathbf{x}_1,...,\mathbf{x}_k) = \prod_{i=1}^{n} & \Big[[f(y_i|x_{1i},...,x_{ki},\theta)]^{\delta_1(i)} \\ & \times \quad [E_{X_1|X_2,...,X_k}[f(y_i|X_{1i},...,x_{ki},\theta)]]^{1-\delta_1(i)}\Big] \end{aligned} \tag{4.9}$$

em que θ é o vetor de parâmetros a ser estimado. Note que, neste caso, precisamos de um modelo para $Y|X_1,...,X_k$ e $X_1|X_2,...,X_k$, mas o processo de estimação é único. Nesse capítulo, consideramos modelos para os quais existe a solução analítica da esperança na Equação (4.9).

Feita a construção para a situação em que apenas uma das variáveis explicativas possui valores faltantes, consideremos, agora, o cenário em que X_1 e X_2 têm informações incompletas.

Observe que, neste caso, para uma amostra de tamanho n de cada uma das variáveis $Y,X_1,...,X_k$, $(y_1,...,y_n,x_{11},...,x_{1n},...,x_{k1},...,x_{kn})$, temos quatro situações possíveis : a) a observação i tem todas as suas variáveis observadas; b) a observação i tem X_1 com valor faltante; c) a observação i tem X_2 com valor faltante; d) a observação i tem X_1 e X_2 com valores faltantes. As situações a) e b) já foram definidas na Equação (4.6) e Equação (4.7), respectivamente, enquanto que a situação c) pode ser obtida de maneira análoga à b), da seguinte forma :

c) Quando X_2 está faltante ou é desconhecida, temos:

$$\begin{aligned} f(y_i|x_{1i},x_{3i},...,x_{ki},\theta) &= \int f(y_i,x_{2i}|x_{1i},x_{3i},...,x_{ki},\theta)dx_{2i} \\ &= \int f(y_i|x_{1i},x_{2i},...,x_{ki},\theta)f(x_{2i}|x_{1i},x_{3i},...,x_{ki},\theta)dx_{2i} \\ &= E_{X_2|X_1,X_3,...,X_k}[f(y_i|x_{1i},X_{2i},...,x_{ki},\theta)] \end{aligned} \tag{4.10}$$

que é a definição da esperança condicional de $f(y_i|x_{1i},X_{2i},...,x_{ki},\theta)$ em relação à $X_2|X_1,X_3,...,X_k$. Observe que, neste caso, precisamos de um modelo para $X_2|X_1,X_3,...,X_k$.

Em relação ao cenário d), como temos valores faltantes em X_1 e X_2, precisamos integrar a função de densidade $f(y,x_{1i},x_{2i}|x_{3i},...,x_{ki},\theta)$ em relação a x_{1i} e x_{2i}. Temos que, dado um modelo para $X_2|X_3,...,X_k$:

$$f(x_{2i}|x_{3i},...,x_{ki},\theta)f(x_{3i},...,x_{ki}|\theta) = f(x_{2i},x_{3i},...,x_{ki}|\theta). \tag{4.11}$$

Além disso, da Equação (4.3) e (4.11), segue que:

$$\begin{aligned} f(y_i,x_{1i},x_{2i}|x_{3i},...,x_{ki},\theta) &= \frac{f(y_i,x_{1i},...,x_{ki}|\theta)}{f(x_{3i},...,x_{ki}|\theta)} \\ &= \frac{f(y_i|x_{1i},...,x_{ki},\theta)f(x_{1i}|x_{2i},...,x_{ki},\theta)f(x_{2i},...,x_{ki}|\theta)}{f(x_{3i},...,x_{ki}|\theta)} \\ &= f(y_i|x_{1i},...,x_{ki},\theta)f(x_{1i}|x_{2i},...,x_{ki},\theta)f(x_{2i}|x_{3i},...,x_{ki},\theta). \end{aligned} \tag{4.12}$$

Logo,

d) Quando X_1 e X_2 estão faltantes ou são desconhecidas, temos:

$$\begin{aligned} f(y_i|x_{3i},...,x_{ki},\theta) &= \int\int f(y_i,x_{1i},x_{2i}|x_{3i},...,x_{ki},\theta)dx_{1i}dx_{2i} \\ &= \int\int f(y_i|x_{1i},...,x_{ki},\theta)f(x_{1i}|x_{2i},...,x_{ki},\theta)f(x_{2i}|x_{3i},...,x_{ki},\theta)dx_{1i}dx_{2i} \\ &= \int\left[\int f(y_i|x_{1i},...,x_{ki},\theta)f(x_{1i}|x_{2i},...,x_{ki},\theta)dx_{1i}\right] f(x_{2i}|x_{3i},...,x_{ki},\theta)dx_{2i} \\ &= \int E_{X_1|X_2,...,X_k}[f(y_i|X_{1i},...,x_{ki},\theta)]f(x_{2i}|x_{3i},...,x_{ki},\theta)dx_{2i} \\ &= E_{X_2|X_3,...,X_k}\left[E_{X_1|X_2,...,X_k}[f(y_i|X_{1i},X_{2i},...,x_{ki},\theta)]\right] \end{aligned} \tag{4.13}$$

que é a definição da esperança em relação a $X_2|X_3,...,X_k$ da esperança condicional de $f(y_i|X_{1i},X_{2i},...,x_{ki},\theta)$ em relação à $X_1|X_2,...,X_k$.

Considerando os indicadores $\delta_1(i)$ e $\delta_2(i)$ relativos às variáveis X_1 e X_2, respectivamente, a função de verossimilhança, então, pode ser escrita como:

$$\begin{aligned} L(\theta|\mathbf{y},\mathbf{x}_1,...,\mathbf{x}_k) = \prod_{i=1}^{n} & \Big[[f(y_i|x_{1i},...,x_{ki},\theta)]^{\delta_1(i)\delta_2(i)} \\ \times\ & [E_{X_1|X_2,...,X_k}[f(y_i|X_{1i},...,x_{ki},\theta)]]^{(1-\delta_1(i))\delta_2(i)} \\ \times\ & [E_{X_2|X_1,X_3,...,X_k}[f(y_i|x_{1i},X_{2i},...,x_{ki},\theta)]]^{(1-\delta_2(i))\delta_1(i)} \\ \times\ & [E_{X_2|X_3,...,X_k}\left[E_{X_1|X_2,...,X_k}[f(y_i|X_{1i},X_{2i},...,x_{ki},\theta)]\right]]^{(1-\delta_1(i))(1-\delta_2(i))}\Big] \end{aligned} \tag{4.14}$$

em que θ é o vetor de parâmetros a ser estimado.

Note que, neste caso, precisamos de um modelo para $Y|X_1,...,X_k, X_1|X_2,...,X_k, X_2|X_1,X_3,...,X_k$ e $X_2|X_3,...,X_k$ mas o processo de estimação é único. Nesse capítulo novamente consideramos modelos em que as esperanças são encontradas analiticamente.

4.2 Modelo Gaussiano com uma variável faltante

Nesta seção, apresentamos um exemplo com resolução analítica da metodologia proposta para lidar com dados faltantes, baseado em um modelo Gaussiano de regressão linear múltipla com k variáveis, em que consideramos, primeiramente, apenas X_1 com valores faltantes e $k = 2$ e, depois, X_1 e X_2 apresentando informações incompletas e $k = 3$. Vale ressaltar que, sempre que assumirmos distribuições normais para a variável resposta e faltantes ou outros casos particulares, a resolução analítica existe.

Sejam Y, X_1, X_2 variáveis aleatórias, ou seja, $k = 2$ e uma amostra de tamanho n para cada uma das variáveis consideradas. Então, para uma observação i, $i = 1, ..., n$, temos:

$$f(y_i|x_{1i}, x_{2i}, \theta) = \frac{f(y_i, x_{1i}, x_{2i}, \theta)}{f(x_{1i}, x_{2i}, \theta)}. \tag{4.15}$$

Sendo assim, como vimos nos cálculos da Seção 4.1:

$$f(y_i, x_{1i}|x_{2i}, \theta) = f(y_i|x_{1i}, x_{2i}, \theta) f(x_{1i}|x_{2i}, \theta). \tag{4.16}$$

Considere, agora, que $f(y_i|x_{1i}, x_{2i}, \theta)$ é a função densidade da distribuição $N\left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \sigma_1^2\right)$ e $f(x_{1i}|x_{2i}, \theta)$ é a função densidade da distribuição $N\left(\gamma_0 + \gamma_1 x_{2i}, \sigma_2^2\right)$. Veja que estamos admitindo, neste caso, que a variável X_1 pode conter dados faltantes.

Logo, para observações com valores faltantes em X_1,

$$\begin{aligned} f(y_i|x_{2i}, \theta) = & \int \left[\frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i})^2}{2\sigma_1^2} \right) \right. \\ & \left. \times \exp\left(-\frac{(x_{1i} - \gamma_0 - \gamma_1 x_{2i})^2}{2\sigma_2^2} \right) \right] dx_{1i}. \end{aligned} \tag{4.17}$$

Integrando a Equação (4.17) em relação a x_{1i}, obtemos:

$$\frac{1}{2\pi\sigma_1\sigma_2} \int \exp\left(-\frac{((y_i - \beta_0 - \beta_2 x_{2i}) - \beta_1 x_{1i})^2}{2\sigma_1^2} \right) \exp\left(-\frac{(x_{1i} - (\gamma_0 + \gamma_1 x_{2i}))^2}{2\sigma_2^2} \right) dx_{1i} \tag{4.18}$$

e fazendo $A = y - \beta_0 - \beta_2 x_{2i}$ e $B = \gamma_0 + \gamma_1 x_{2i}$, a Equação (4.18) se torna:

$$\begin{aligned} & \frac{1}{2\pi\sigma_1\sigma_2} \int \exp\left(-\frac{(A - \beta_1 x_{1i})^2}{2\sigma_1^2} \right) \exp\left(-\frac{(x_{1i} - B)^2}{2\sigma_2^2} \right) dx_{1i} \\ = & \frac{1}{2\pi\sigma_1\sigma_2} \int \exp\left(\frac{-(\beta_1^2\sigma_2^2 + \sigma_1^2)x_{1i}^2 + 2(A\beta_1\sigma_2^2 + B\sigma_1^2)x_{1i} - (A^2\sigma_2^2 + B^2\sigma_1^2)}{2\sigma_1^2\sigma_2^2} \right) dx_{1i}. \end{aligned}$$

Fazendo $C = \beta_1^2\sigma_2^2 + \sigma_1^2$, $D = A\beta_1\sigma_2^2 + B\sigma_1^2$ e $E = A^2\sigma_2^2 + B^2\sigma_1^2$, temos:

$$
\begin{aligned}
& \frac{1}{2\pi\sigma_1\sigma_2}\int \exp\left(\frac{-Cx_{1i}^2 + 2Dx_{1i} - E}{2\sigma_1^2\sigma_2^2}\right) dx_{1i} \\
= \; & \frac{1}{2\pi\sigma_1\sigma_2}\exp\left(\frac{-E}{2\sigma_1^2\sigma_2^2}\right)\int \exp\left(\frac{-Cx_{1i}^2 + 2Dx_{1i}}{2\sigma_1^2\sigma_2^2}\right) dx_{1i} \\
= \; & \frac{1}{2\pi\sigma_1\sigma_2}\exp\left(\frac{-E}{2\sigma_1^2\sigma_2^2}\right)\exp\left(\frac{D^2}{2C\sigma_1^2\sigma_2^2}\right)\int \exp\left(-\frac{1}{2}\left(\frac{x_{1i} - \frac{D}{C}}{\frac{\sigma_1\sigma_2}{\sqrt{C}}}\right)^2\right) dx_{1i} \\
= \; & \frac{1}{2\pi\sigma_1\sigma_2}\exp\left(\frac{-E}{2\sigma_1^2\sigma_2^2}\right)\exp\left(\frac{D^2}{2C\sigma_1^2\sigma_2^2}\right)\frac{\sqrt{2\pi}\sigma_1\sigma_2}{\sqrt{C}} \\
= \; & \frac{1}{\sqrt{2\pi C}}\exp\left(\frac{-EC + D^2}{2C\sigma_1^2\sigma_2^2}\right) \\
= \; & \frac{1}{\sqrt{2\pi C}}\exp\left(-\frac{1}{2}\left(\frac{(A - B\beta_1)^2}{C}\right)\right) \\
= \; & \frac{1}{\sqrt{2\pi(\beta_1^2\sigma_2^2 + \sigma_1^2)}}\exp\left(-\frac{1}{2}\left(\frac{(y_i - \beta_0 - \beta_2 x_{2i} - (\gamma_0 + \gamma_1 x_{2i})\beta_1)^2}{\beta_1^2\sigma_2^2 + \sigma_1^2}\right)\right) \\
= \; & \frac{1}{\sqrt{2\pi(\beta_1^2\sigma_2^2 + \sigma_1^2)}}\exp\left(-\frac{1}{2}\left(\frac{(y_i - (\beta_0 + \beta_2 x_{2i} + (\gamma_0 + \gamma_1 x_{2i})\beta_1)^2)}{\beta_1^2\sigma_2^2 + \sigma_1^2}\right)\right). \qquad (4.19)
\end{aligned}
$$

Podemos construir a função de verossimilhança deste modelo, com base no que foi discutido na Seção 4.1, da seguinte forma:

$$
L(\theta|\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2) = \prod_{i=1}^{n} \left[f(y_i|x_{1i}, x_{2i}, \theta)^{\delta_1(i)} \times f(y_i|x_{2i}, \theta)^{1-\delta_1(i)} \right]. \qquad (4.20)
$$

Observe que o espaço paramétrico θ da função de verossimilhança expressa pela Equação 4.20 para este caso em que temos uma regressão linear múltipla com duas variáveis explicativas, sendo uma delas com valores faltantes, se torna $\theta = (\beta_0, \beta_1, \beta_2, \gamma_0, \gamma_1, \sigma_1^2, \sigma_2^2)$.

Para obtermos a função de log-verossimilhança $l(\theta|\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2)$, basta fazermos $\log L(\theta|\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2)$:

$$
\begin{aligned}
l(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2) &= \log L(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2) \\
&= \log\left[\prod_{i=1}^{n}\left[f(y_i|x_{1i},x_{2i},\theta)^{\delta_1(i)}\times f(y_i|x_{2i},\theta)^{1-\delta_1(i)}\right]\right] \\
&= \sum_{i=1}^{n}\left[\delta_1(i)\log f(y_i|x_{1i},x_{2i},\theta)+(1-\delta_1(i))\log f(y_i|x_{2i},\theta)\right] \\
&= \sum_{i=1}^{n}\left[\delta_1(i)\log\left[\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{1}{2}\left(\frac{(y_i-(\beta_0+\beta_1x_{1i}+\beta_2x_{2i}))^2}{\sigma_1^2}\right)\right)\right]\right. \\
&+ (1-\delta_1(i))\log\left[\frac{1}{\sqrt{2\pi(\beta_1^2\sigma_2^2+\sigma_1^2)}}\right. \\
&\times \left.\left.\exp\left(-\frac{1}{2}\left(\frac{(y_i-(\beta_0+\beta_2x_{2i}+(\gamma_0+\gamma_1x_{2i})\beta_1))^2}{\beta_1^2\sigma_2^2+\sigma_1^2}\right)\right)\right]\right]. \qquad (4.21)
\end{aligned}
$$

Para encontrarmos as estimativas dos parâmetros, temos que maximizar a função de log-verossimilhança dada pela Equação (4.21).

4.2.1 Análise preditiva do método

Para analisarmos o método proposto em relação à sua capacidade preditiva, dividimos o banco de dados em 70% para treino (subconjunto através do qual estimamos os parâmetros) e 30% para teste (subconjunto para o qual calculamos os valores preditos $\widehat{y}$ e comparamos com os observados y). Para o cálculo do $\widehat{y}$ consideramos o valor esperado estimado da distribuição de $Y|X_1,X_2$ para observações completas ou de $Y|X_2$, caso a observação i possua valor faltante em relação a X_1. Logo,

i) Se x_{1i} é observado:

$$\widehat{y_i} = \widehat{\beta_0}+\widehat{\beta_1}x_{1i}+\widehat{\beta_2}x_{2i}, \qquad (4.22)$$

em que $\widehat{\beta_0}$, $\widehat{\beta_1}$ e $\widehat{\beta_2}$ são os valores das estimativas dos parâmetros e, como $f(y_i|x_{1i},x_{2i},\theta)$ é a função densidade da distribuição $N\left(\beta_0+\beta_1x_{1i}+\beta_2x_{2i},\sigma_1^2\right)$, então $\widehat{y_i}$ dado pela Equação (4.22) se refere ao valor esperado estimado da distribuição de $Y|X_1,X_2$;

ii) Se x_{1i} é faltante:

$$\widehat{y_i} = \widehat{\beta_0}+\widehat{\beta_2}x_{2i}+(\widehat{\gamma_0}+\widehat{\gamma_1}x_{2i})\widehat{\beta_1}, \qquad (4.23)$$

em que $\widehat{\beta_0}$, $\widehat{\beta_1}$, $\widehat{\beta_2}$, $\widehat{\gamma_0}$ e $\widehat{\gamma_1}$ são os valores das estimativas dos parâmetros e, como $f(y_i|x_{2i},\theta)$ é a função densidade da distribuição $N\left(\beta_0+\beta_2x_{2i}+(\gamma_0+\gamma_1x_{2i})\beta_1,\beta_1^2\sigma_2^2+\sigma_1^2\right)$, então $\widehat{y_i}$ dado pela Equação (4.23) se refere ao valor esperado estimado da distribuição de $Y|X_2$.

Efetuado o cálculo do $\widehat{y}$ para todas as observações da base teste, calculamos a média das diferenças quadráticas entre $\widehat{y}$ e y, o erro quadrático médio da predição. Quanto mais próximo de zero estiver essa soma quadrática, melhor é a predição do modelo estimado.

4.2.2 Estudo de simulação

Nesta seção, será discutido como foi feito o estudo de simulação e comparação do desempenho do método proposto com o de outros métodos de imputação e deleção de dados, considerando o modelo Gaussiano. Nesse cenário com apenas uma variável com valores faltantes, comparamos o desempenho dos diferentes métodos em relação ao viés e ao erro quadrático médio (EQM) dos valores das estimativas dos parâmetros e em relação ao poder preditivo definido pela média das diferenças quadráticas entre $\widehat{y}$ e y em amostras teste, separadas especificamente para esse fim.

Vários cenários de simulação foram testados, entre os quais variamos: o tamanho da amostra ($n = 100$ e $n = 300$), a proporção de valores faltantes ($p = 0.20$ e $p = 0.60$) e o mecanismo que gera os dados faltantes (MCAR, MAR e MNAR). Para cada cenário analisado, simulamos 30 réplicas (amostras diferentes) sob as mesmas condições. Com elas, temos amostras de tamanho 30 para conduzir análises de desempenho através do viés e EQM das estimativas dos parâmetros, assim como do erro de predição. Quanto mais próximas de zero essas diferenças, mais precisa é a estimação e a predição. Aqui, adotamos o modelo normal como a distribuição de probabilidades das variáveis, mas qualquer outro modelo que possibilite a resolução analítica pode ser considerado e a metodologia adaptada.

Para cada conjunto de dados, separamos as 70% primeiras observações para treino e estimação, através da qual fazemos a análise inferencial dos parâmetros e os outros 30% para teste, em que analisamos o método de estimação baseado em modelo na presença de valores faltantes em relação ao poder preditivo, comparando-o com os métodos: modelo completo (sem dados faltantes), método de deleção *listwise*, método de imputação pela média, método de imputação por *Random Forest*, método de imputação por *hot deck* e método de imputação múltipla.

Realizamos as simulações de acordo com os seguintes passos:

Passo 1: Geramos x_{2i} da distribuição Normal com média 0 e variância 1 e, em seguida, x_{1i} da distribuição Normal de média $\gamma_0 + \gamma_1 x_{2i}$ e variância σ_2^2 e y_i da distribuição Normal de média $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$ e variância σ_1^2, para $i = 1, ..., n$, sendo n, portanto, o tamanho da amostra final. Os verdadeiros valores destes parâmetros considerados são: $\sigma_1^2 = 2,\ \sigma_2^2 = 4,\ \beta_0 = -1,\ \beta_1 = 1.5,\ \beta_2 = -0.75,\ \gamma_0 = 1$ e $\gamma_1 = 1.5$;

Passo 2: De acordo com o mecanismo considerado e o valor de $p,\ 0 \leq p \leq 1$, geramos valores faltantes na variável X_1 da seguinte forma:

a) Se o mecanismo é MCAR: geramos n valores da distribuição Uniforme entre 0 e 1 e, se o valor na posição i, $i \leq n$, for menor ou igual a proporção p que queremos obter de valores faltantes, então, o i-ésimo valor da variável X_1 será um valor faltante e substituído por NA;

b) Se o mecanismo é MAR: identificamos os $p \times n$ maiores valores da variável X_2 e substituímos por NA os valores de X_1 das observações correspondentes a estes casos;

c) Se o mecanismo é MNAR: identificamos os $p \times n$ maiores valores da variável X_1 e os substituímos por NA.

Ressaltamos que no caso de estimação pelo modelo completo, para comparação de desempenho, esse passo não é realizado;

Passo 3: Maximizamos as funções de log-verossimilhança utilizando o *optim* do software estatístico R com parâmetro $fnscale = -1$. O método numérico usado para encontrar o máximo das funções de log-verossimilhança é o de Nelder-Mead, que é comumente aplicado em problemas de otimização não-linear para os quais as derivadas não podem ser encontradas ou não são definidas. Para obtermos melhores estimativas dos parâmetros, fazemos uma primeira maximização usando como valores iniciais valores não informativos e, em seguida, maximizamos as funções de log-verossimilhança novamente, também utlizando o método de Nelder-Mead e considerando, como valores iniciais, os valores das estimativas obtidos pelo primeiro processo de maximização. Quanto às funções de log-verossimilhança maximizadas, temos que:

a) Para o método proposto nesse trabalho, a função de log-verossimilhança é a construída na Seção 4.2, em que consideramos os dois cenários possíveis para cada observação do conjunto de treinamento ou estimação, ou seja, a observação i ser observada em relação à variável X_1 ou a observação i ser faltante em X_1;

b) Para o método de imputação pela média, primeiramente calculamos a média dos valores do conjunto de treinamento correspondentes à variável X_1 que não estão faltantes e, em seguida, para as observações que possuem valores faltantes em X_1, imputamos esta média. Após, com o conjunto de dados completo obtido com esta imputação, maximizamos a função de log-verossimilhança da Seção 4.2 referente ao cenário em que as observações são completas em relação a X_1;

c) Para o método de imputação por *Random Forest*, imputamos os valores faltantes em relação a X_1 utilizando o pacote *missForest* (STEKHOVEN, 2011). Esta função é usada para imputar valores ausentes usando uma floresta aleatória treinada nos valores observados para prevê-los. Após, com o conjunto de dados completo obtido com esta imputação, maximizamos a função de log-verossimilhança da Seção 4.2 referente ao cenário em que as observações são completas em relação a X_1;

d) Para o método de imputação por *Hot-Deck*, imputamos os valores faltantes em relação a X_1 utilizando o pacote *VIM* (TEMPL *et al.*, 2021). A função *hotdeck* neste pacote é usada para imputar valores ausentes usando o algoritmo hot-deck sequencial e aleatório dentro do domínio da variável faltante. Após, com o conjunto de dados completo obtido com esta imputação, maximizamos a função de log-verossimilhança da Seção 4.2 referente ao cenário em que as observações são completas em relação a X_1;

e) Para o método de imputação múltipla, imputamos os valores faltantes em relação a X_1 utilizando o pacote *Amelia* (HONAKER; KING; BLACKWELL, 2021). Esse pacote realiza a imputação múltipla de dados incompletos multivariados através da combinação da técnica *bootstrap* e algoritmo EM para prever os dados faltantes. O método inclui transformações de normalização, priorização em nível de célula e métodos para lidar com dados transversais de séries temporais. Por meio deste pacote, criamos cinco conjuntos de dados completos e, para encontrarmos a estimativa final dos parâmetros, calculamos a média aritmética das cinco estimativas para cada conjunto de dados completos encontradas por meio da maximização da função de log-verossimilhança da Seção 4.2 referente ao cenário em que as observações são completas em relação a X_1;

f) Para o método de deleção de casos, foram removidas todas as observações que tinham valores faltantes em X_1 e então, com a subamostra de valores completos restante, maximizamos a função de log-verossimilhança da Seção 4.2 referente a este cenário;

g) Por fim, consideramos o caso em que não temos valores faltantes, utilizamos o conjunto de dados completo para maximizarmos a função de log-verossimilhança da Seção 4.2 referente a este cenário.

Passo 4: Após obtermos os valores das estimativas dos parâmetros para cada conjunto de dados dentro de cada método, calculamos a diferença e a diferença quadrática dessas estimativas em relação aos verdadeiros valores dos parâmetros. Como os parâmetros γ_0, γ_1 e σ_2^2 são estimados diretamente apenas pela metodologia proposta, os índices de desempenho dos seus estimadores não são mostrados e comparados;

Passo 5: Para analisarmos os métodos em relação ao poder preditivo, calculamos o erro quadrático médio do $\widehat{y}_i$ em relação ao y_i observado para todas as observações do conjunto de teste. O cálculo de $\widehat{y}_i$ se dá da seguinte forma:

a) Para a metodologia proposta e cada observação i do conjunto de dados de teste, calculamos $\widehat{y}_i$ de acordo com o proposto na Seção 4.2.1;

b) Para os métodos de imputação de dados por *Random Forest*, imputação por *Hot-Deck* e imputação múltipla, realizamos a imputação dos dados faltantes na base de teste usando os mesmos procedimentos do passo 3 para cada método e, com os dados

completos, calculamos $\widehat{y}_i$ de acordo com o item i) da Sessão 4.2.1. Para o caso da imputação múltipla, como criamos cinco conjuntos de dados completos, o erro quadrático médio é dado pela média entre os erros quadráticos médios dos cinco conjuntos gerados;

c) Para o método de imputação pela média, a média das observações não faltantes em X_1 do conjunto de treino é o valor utilizado para ser imputado nas observações que possuem valores faltantes no conjunto de teste e, então, após esta imputação, calculamos $\widehat{y}_i$ de acordo com o item i) da Sessão 4.2.1;

d) Para o método em que consideramos o conjunto de dados de teste completo, sem valores faltantes, calculamos $\widehat{y}_i$ de acordo com o item i) da Sessão 4.2.1;

e) Para o método de deleção de dados, como deletamos as observações que possuem valores faltantes, não conseguimos calcular o $\widehat{y}_i$ associado a elas pois não existe, de fato, um processo de imputação ou predição de valores faltantes. Logo, não analisamos o desempenho de predição desse método nas observações do conjunto de teste, por não fazer sentido essa comparação.

As Figuras 1, 3, 5 e 7 mostram o desempenho inferencial do métodos comparados para os 4 cenários analisados com mecanismo MCAR. Em vez de calcularmos um único valor de viés e EQM para a estimativa de cada parâmetro, como sendo a média das diferenças (viés) e das diferenças quadráticas (EQM) observadas entre estimativa e parâmetro nas 30 réplicas, preferimos exibir todas diferenças observadas através de boxplots, para melhor comparação. As Figuras 2, 4, 6 e 8 mostram os erros quadráticos médios de predição observados para as mesmas simulações.

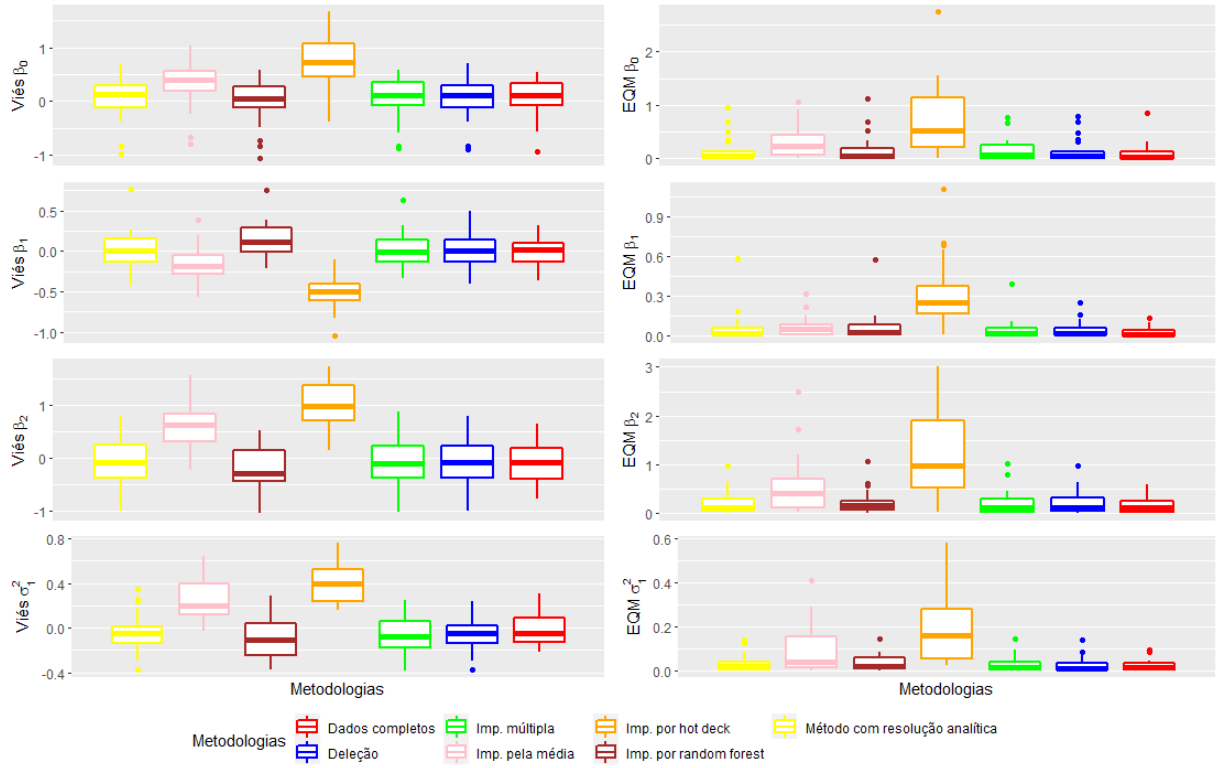

Figura 1 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário MCAR, $n = 100$ e $p = 0.20$.

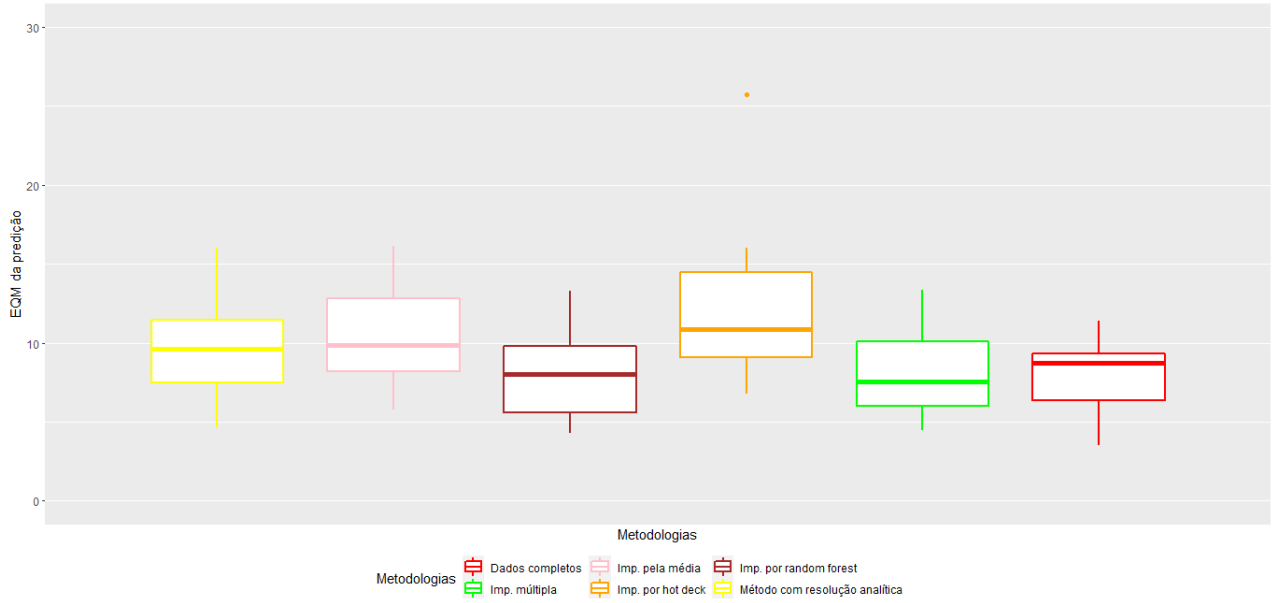

Figura 2 – Erro quadrático médio dos valores preditos para o cenário MCAR, $n = 100$ e $p = 0.20$.

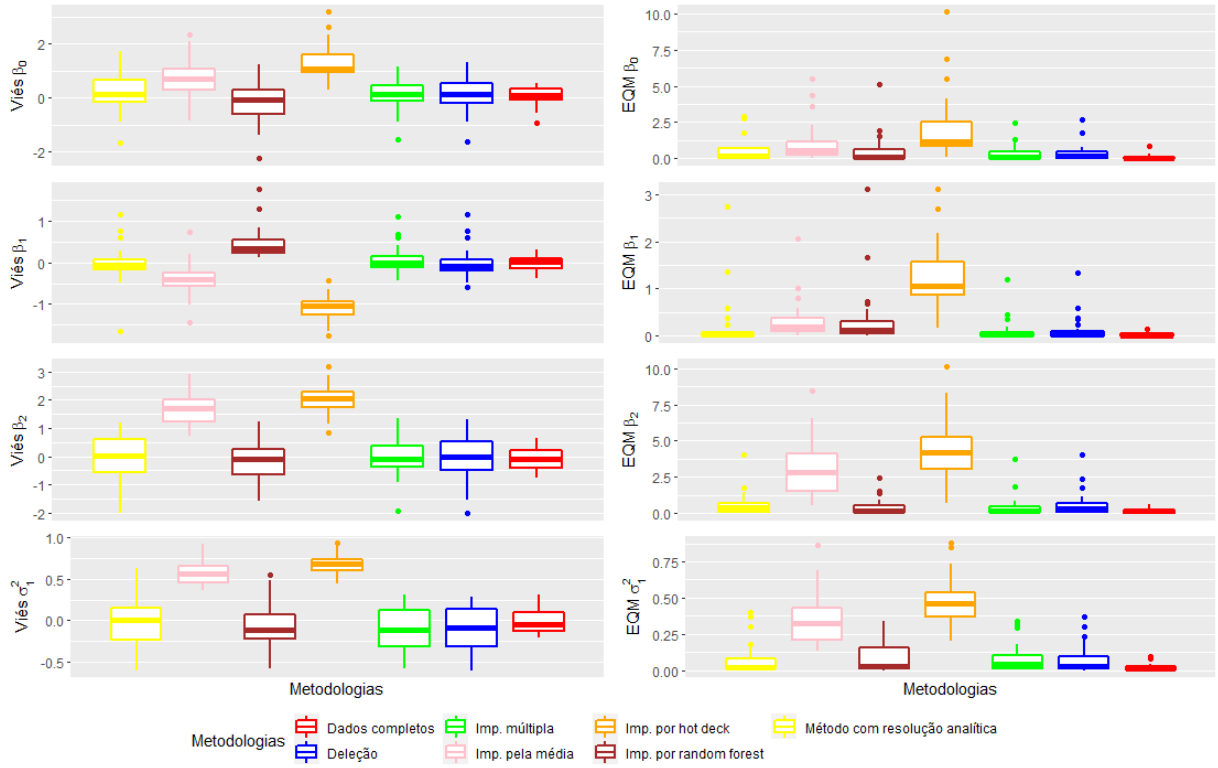

Figura 3 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário MCAR, $n = 100$ e $p = 0.60$.

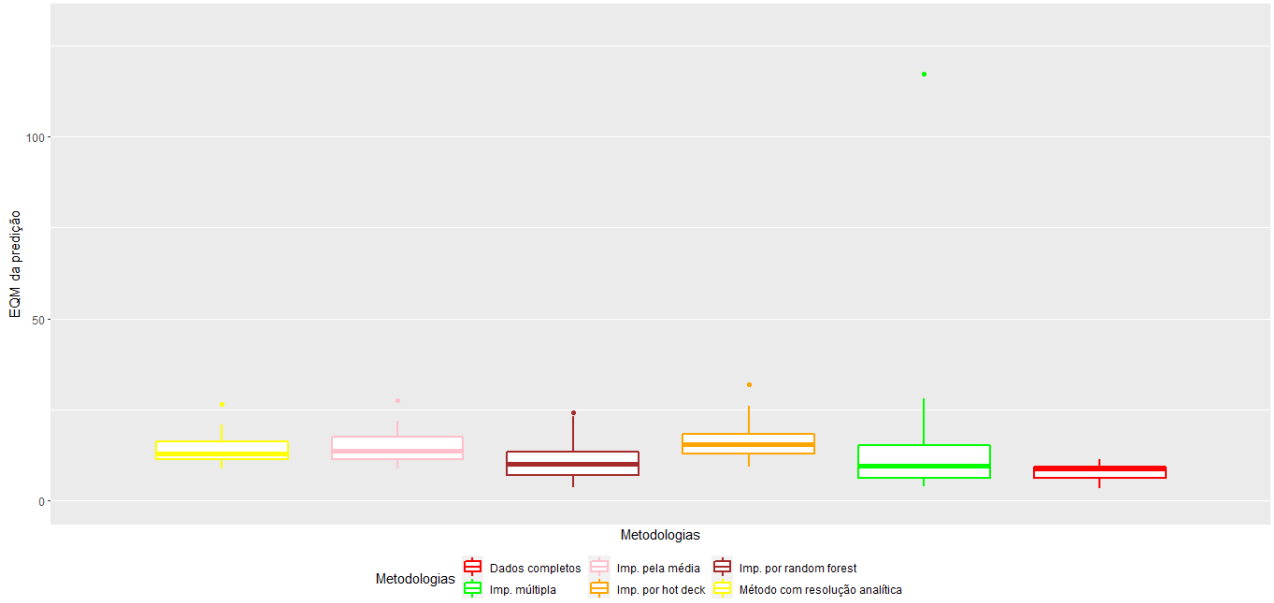

Figura 4 – Erro quadrático médio dos valores preditos para o cenário MCAR, $n = 100$ e $p = 0.60$.

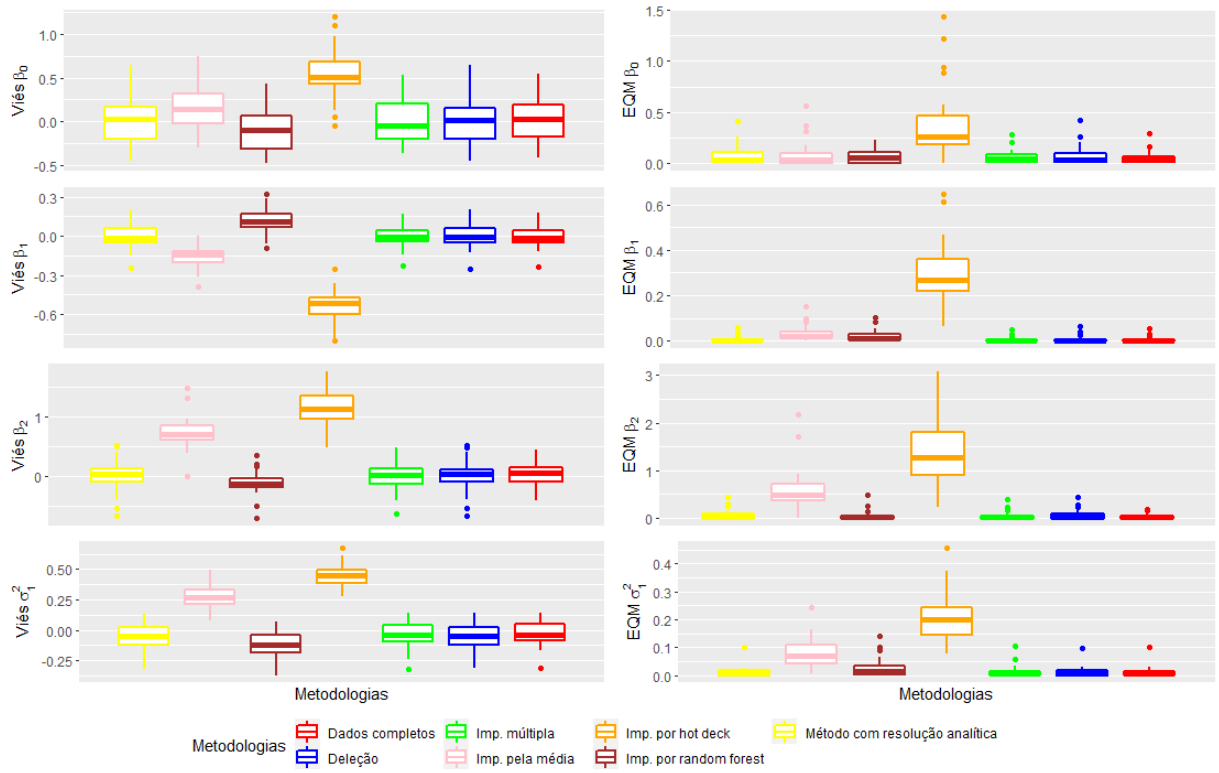

Figura 5 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário MCAR, $n = 300$ e $p = 0.20$.

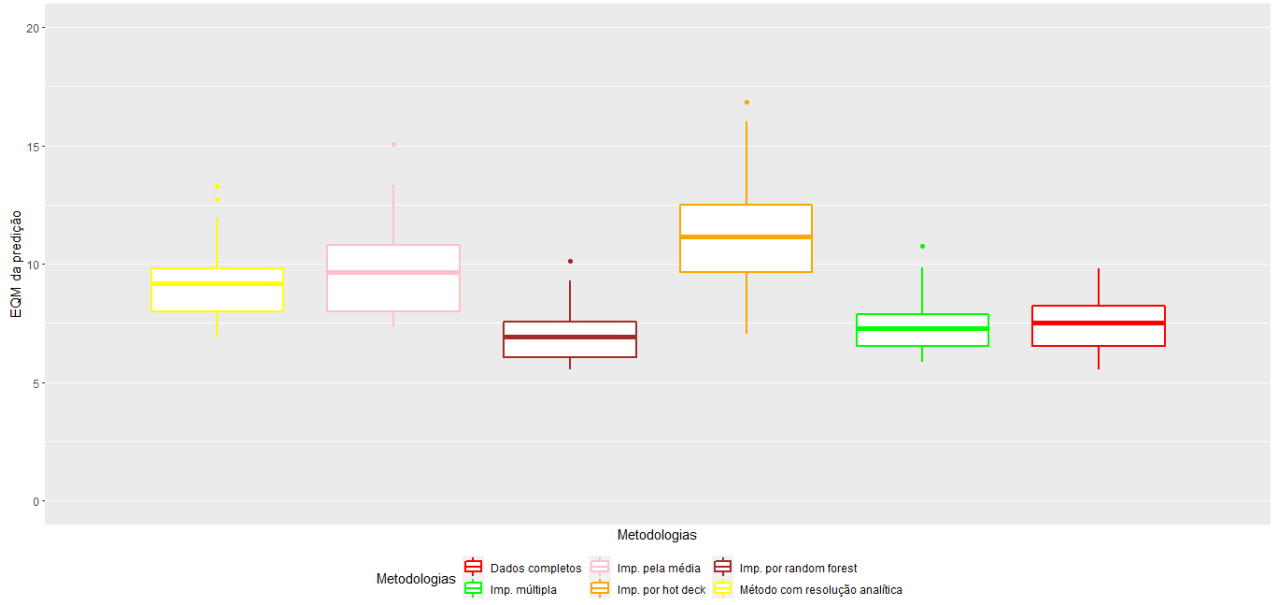

Figura 6 – Erro quadrático médio dos valores preditos para o cenário MCAR, $n = 300$ e $p = 0.20$.

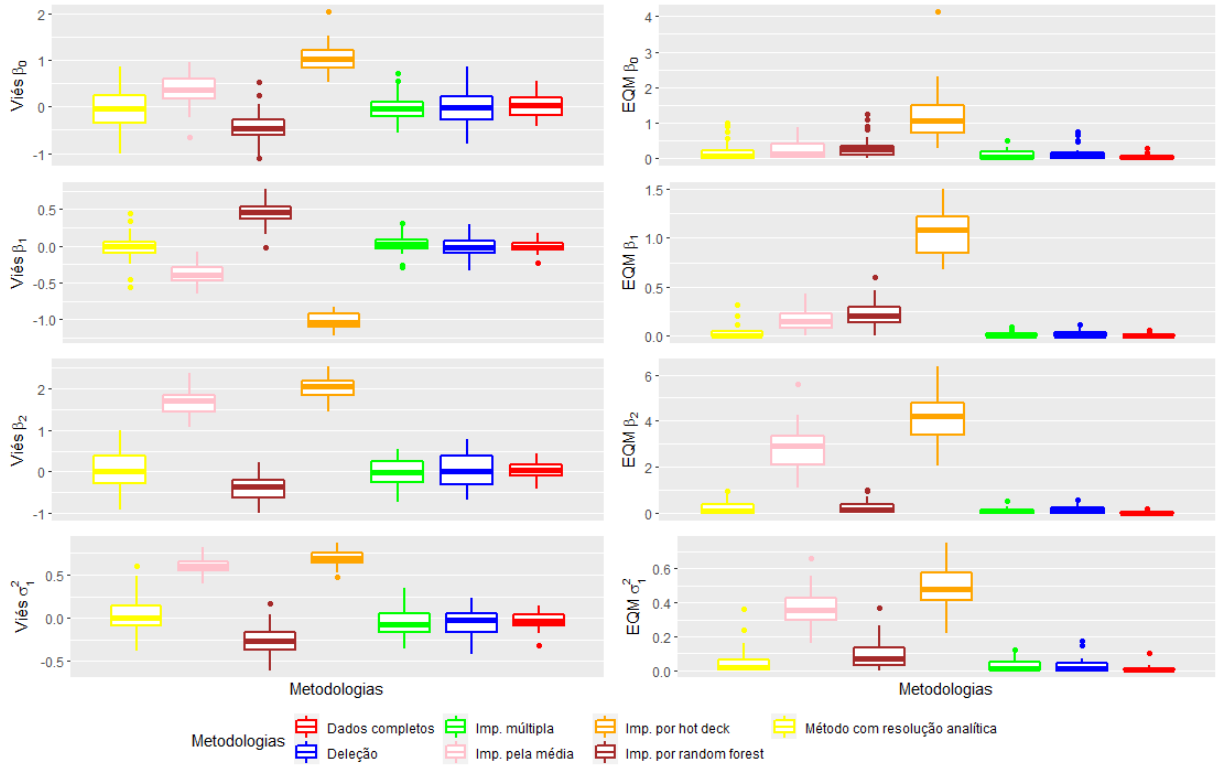

Figura 7 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário MCAR, $n = 300$ e $p = 0.60$.

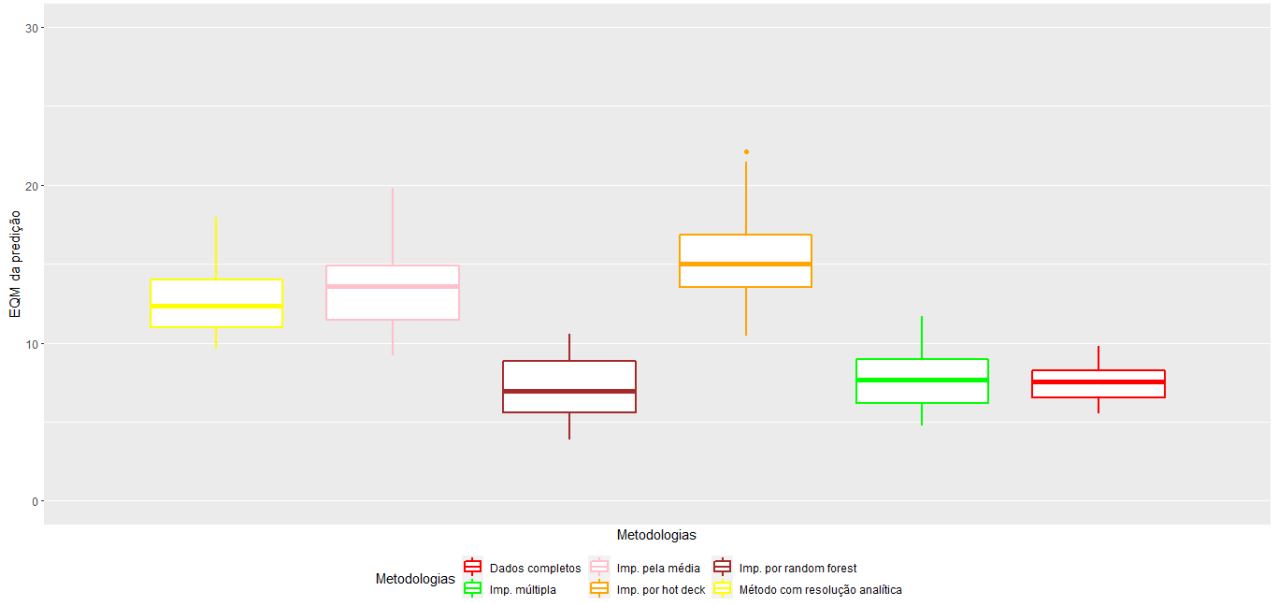

Figura 8 – Erro quadrático médio dos valores preditos para o cenário MCAR, $n = 300$ e $p = 0.60$.

Considerando o mecanismo MCAR de geração de dados faltantes, observamos que a metodologia proposta apresenta resultados inferenciais muito parecidos à análise dos dados completos ou sob delação de dados faltantes nos 4 cenários considerados. Isso representa uma ótima performance do método na estimação de parâmetros na presença de dados faltantes. O método de imputação múltipla também apresenta resultados parecidos, enquanto que os métodos por imputação por *Hot-deck*, por média e *Random Forest* apresentam, nessa ordem, as piores performances.

Em relação ao desempenho preditivo nas amostras de teste, os métodos de imputação múltipla e *Random Forest* se sobressaem ao método proposto. Esse comportamento faz sentido pois essas metodologias são focadas na predição e o mecanismo de geração dos dados faltantes é o completamente aleatório, quando é esperado que todas as metodologias performem de maneira razoável. O método baseado em modelos com resolução analítica, em contrapartida, contempla a estimação de 3 parâmetros adicionais (γ_0, γ_1 e σ_2^2) que propicia um maior acúmulo de erro na predição desse mecanismo de dados faltantes mais simples.

As Figuras 9, 11, 13 e 15 apresentam os resultados inferenciais dos métodos para o mecanismo MAR de dados faltantes e as Figuras 10, 12, 14 e 16 o desempenho preditivo.

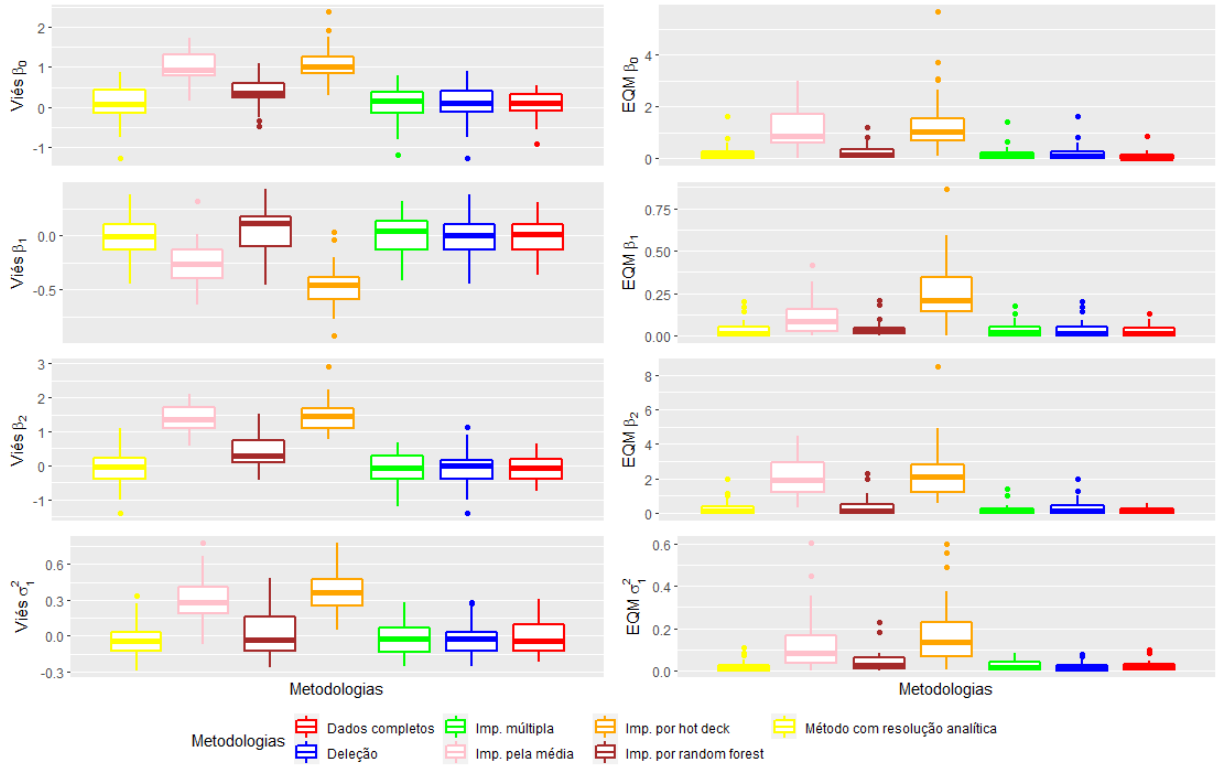

Figura 9 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário MAR, $n = 100$ e $p = 0.20$.

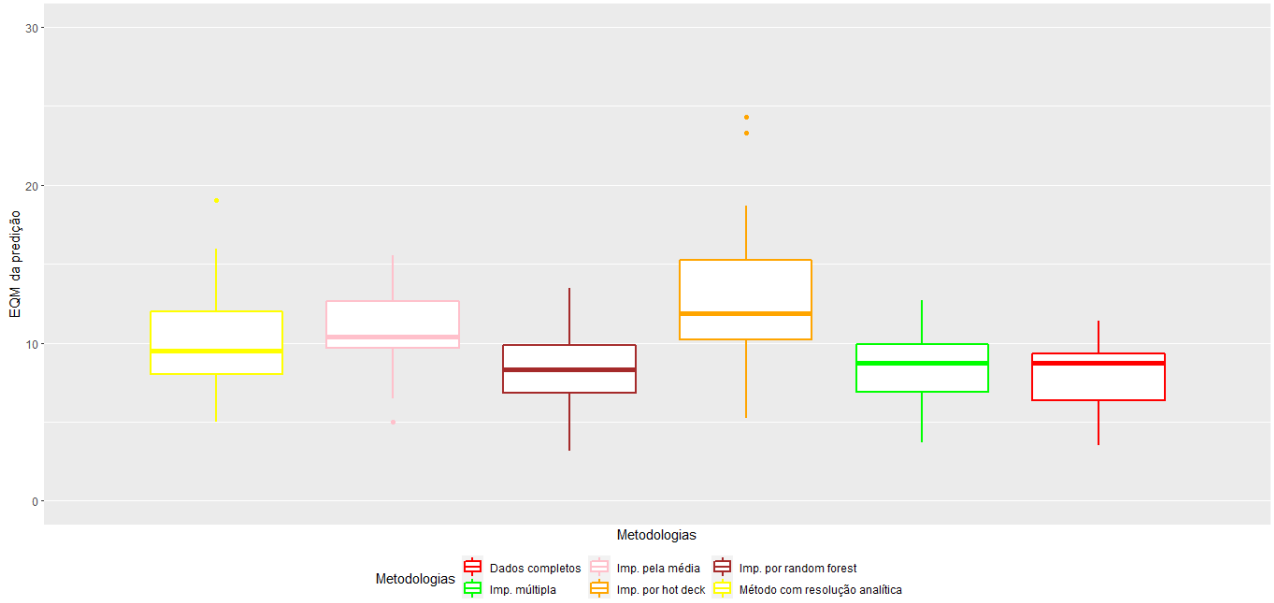

Figura 10 – Erro quadrático médio dos valores preditos para o cenário MAR, $n = 100$ e $p = 0.20$.

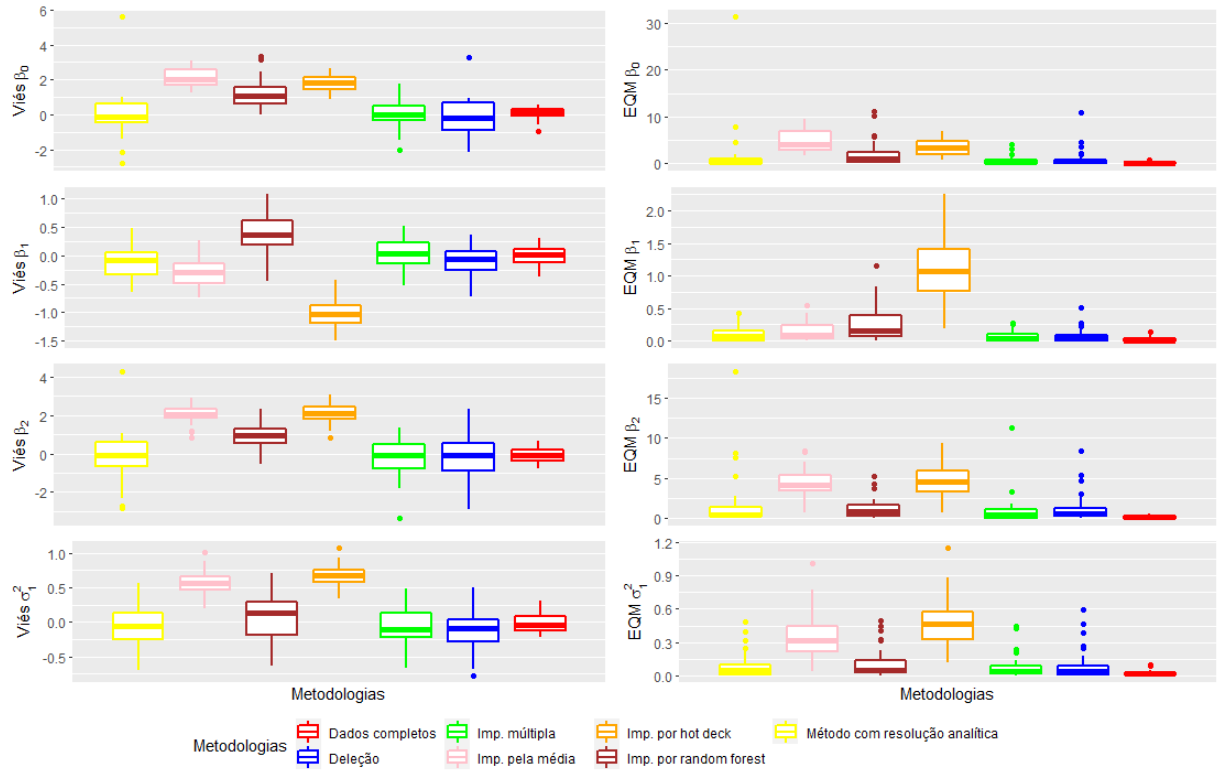

Figura 11 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário MAR, $n = 100$ e $p = 0.60$.

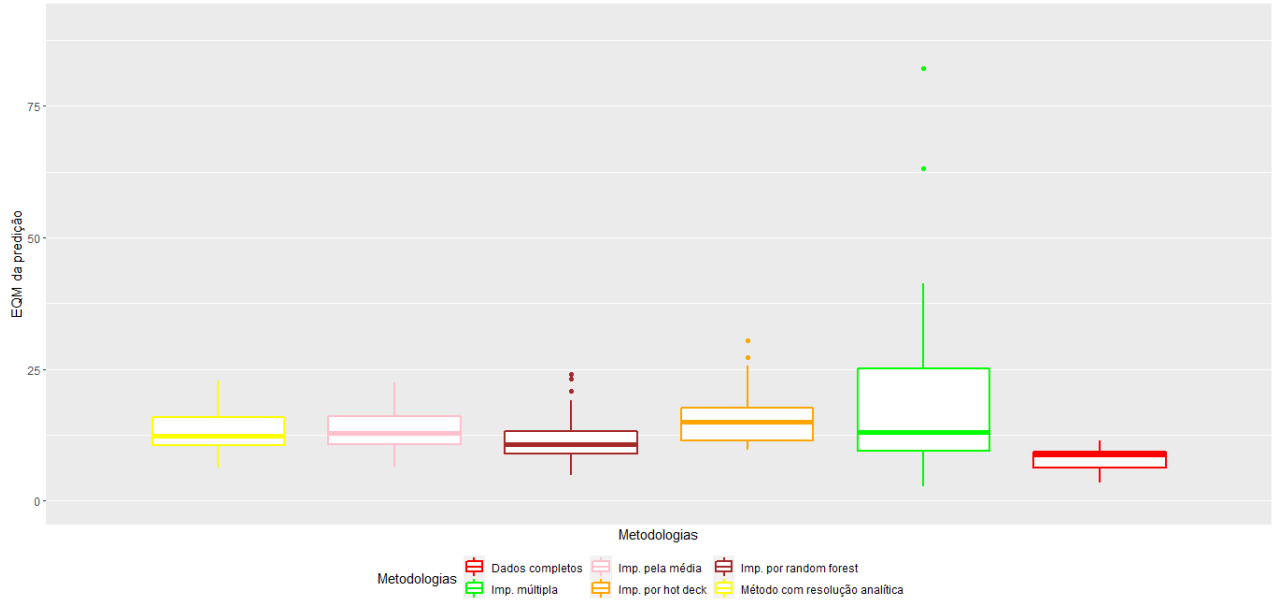

Figura 12 – Erro quadrático médio dos valores preditos para o cenário MAR, $n = 100$ e $p = 0.60$.

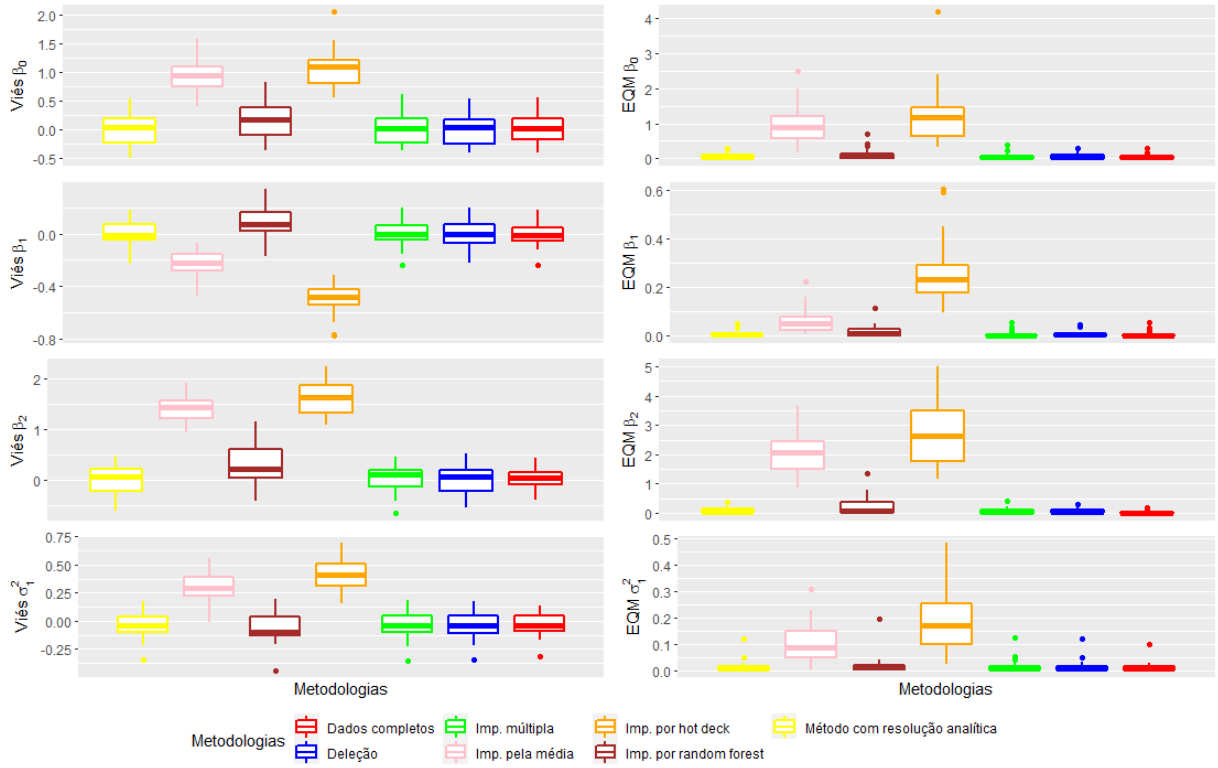

Figura 13 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário MAR, $n = 300$ e $p = 0.20$.

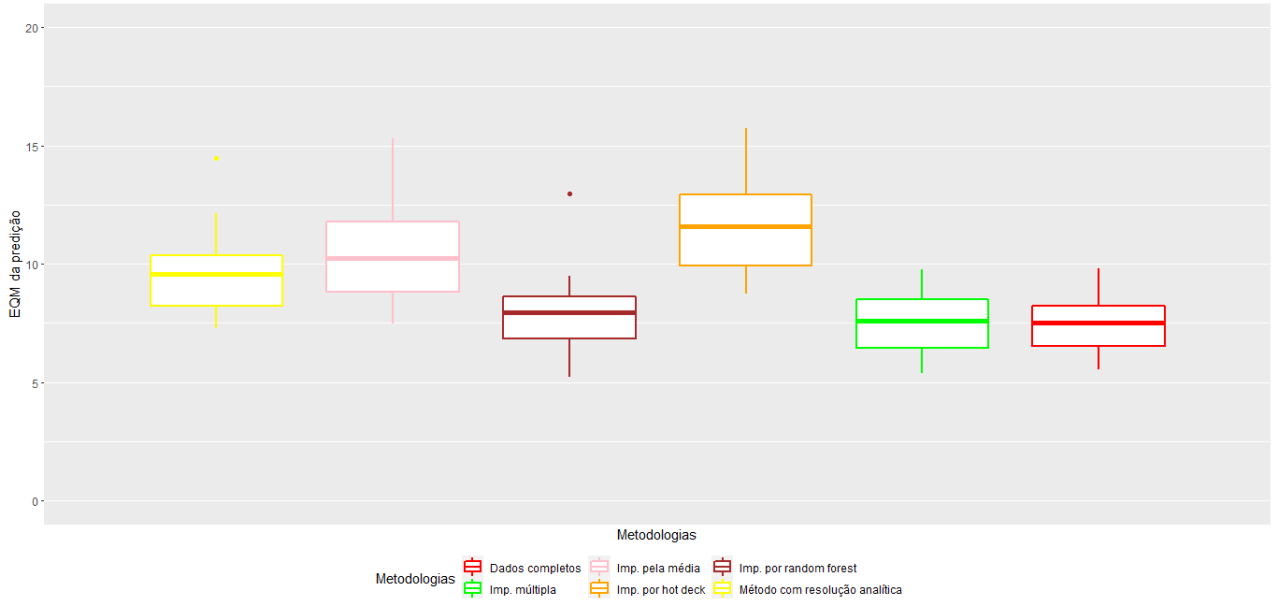

Figura 14 – Erro quadrático médio dos valores preditos para o cenário MAR, $n = 300$ e $p = 0.20$.

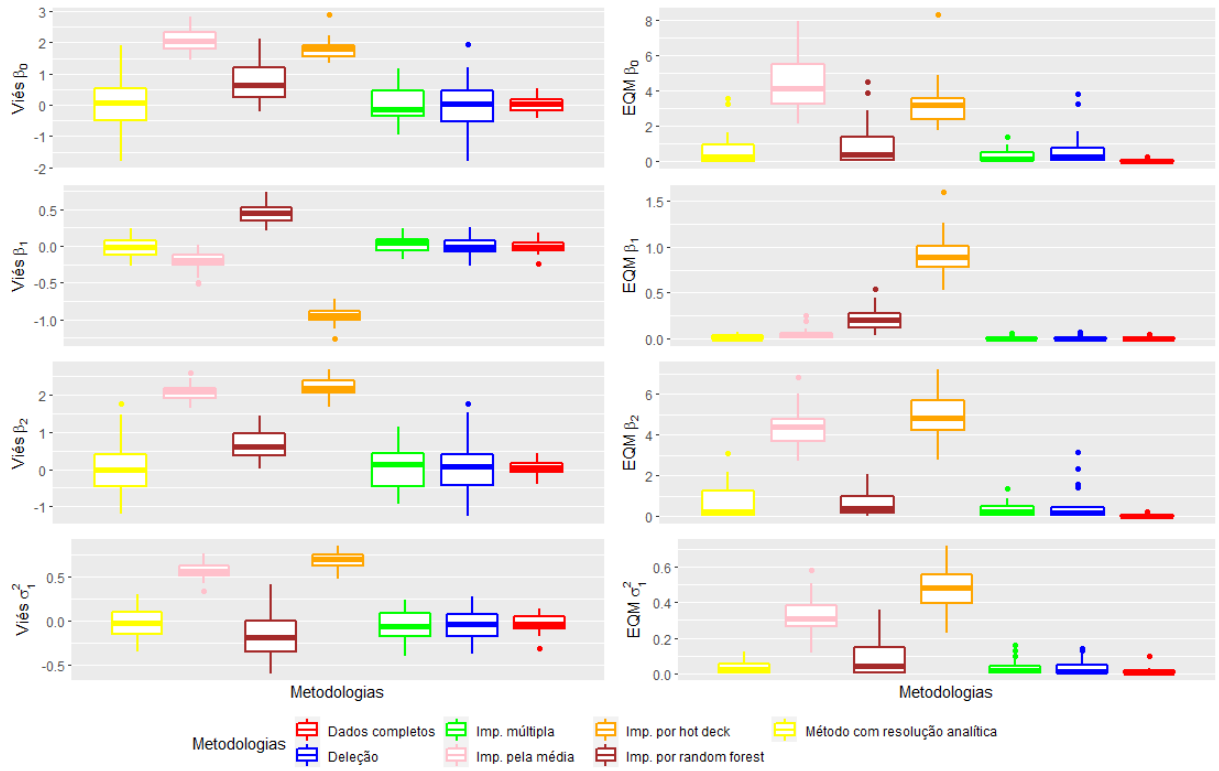

Figura 15 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário MAR, $n = 300$ e $p = 0.60$.

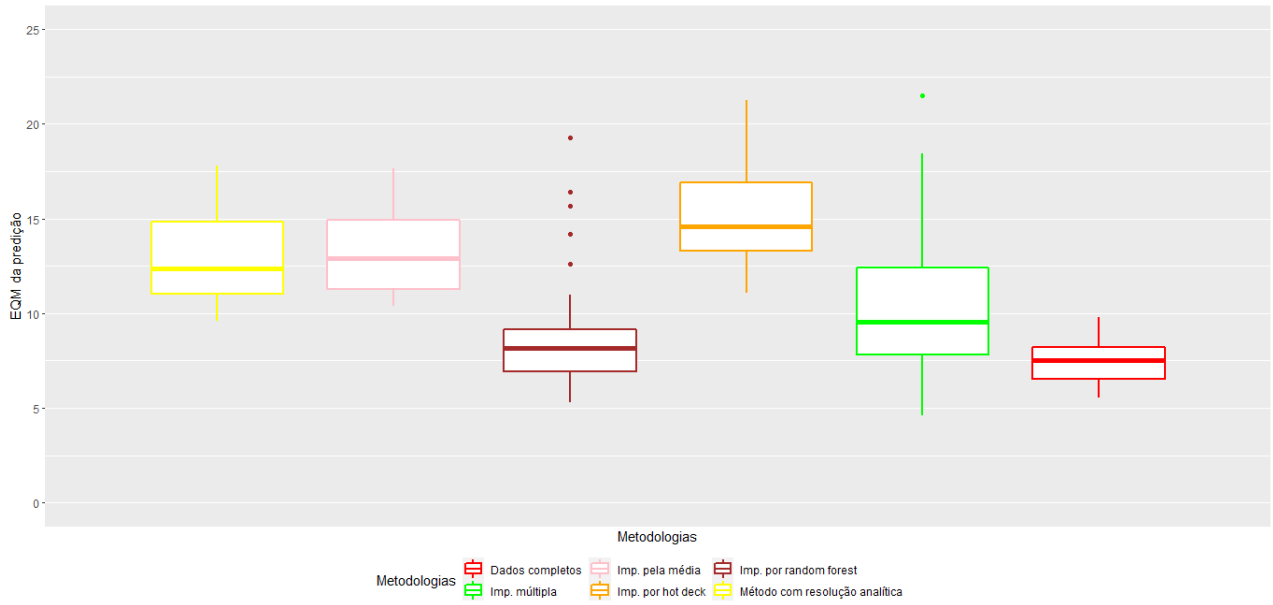

Figura 16 – Erro quadrático médio dos valores preditos para o cenário MAR, $n = 300$ e $p = 0.60$.

Para o mecanismo de geração dos dados faltantes MAR, observamos resultados de estimação muito parecidos aos obtidos no método MCAR. O desempenho preditivo da metodologia proposta, por sua vez, se aproxima ao desempenho da imputação múltipla e do *Random Forest* e, nos cenários de amostras menores, apresenta melhor desempenho ou desempenho semelhante à imputação múltipla.

As Figuras 17, 19, 21 e 23 apresentam os resultados inferenciais dos métodos para o mecanismo MNAR de dados faltantes e as Figuras 18, 20, 22 e 24 o desempenho preditivo.

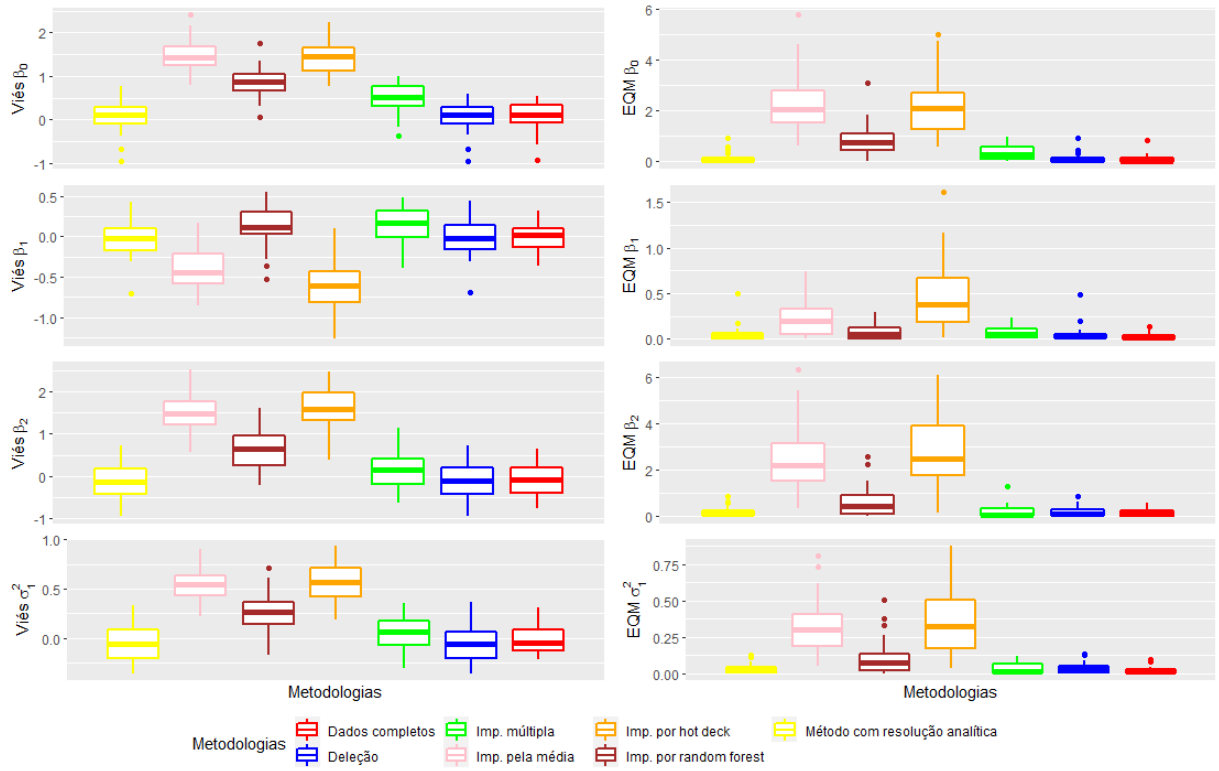

Figura 17 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário MNAR, $n = 100$ e $p = 0.20$.

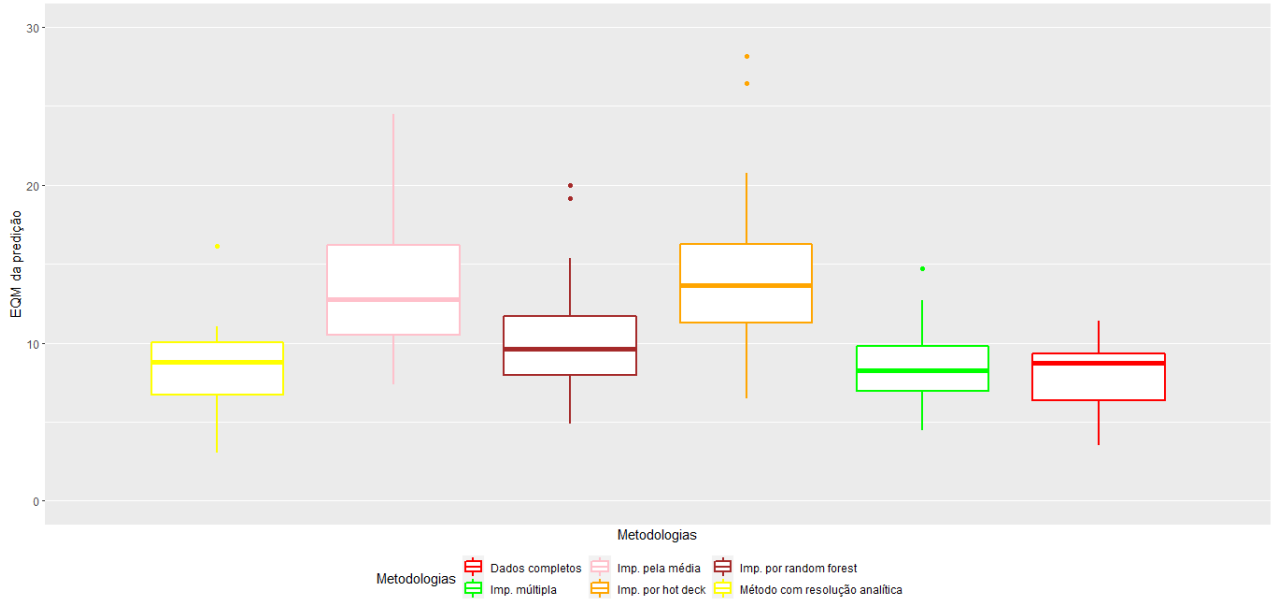

Figura 18 – Erro quadrático médio dos valores preditos para o cenário MNAR, $n = 100$ e $p = 0.20$.

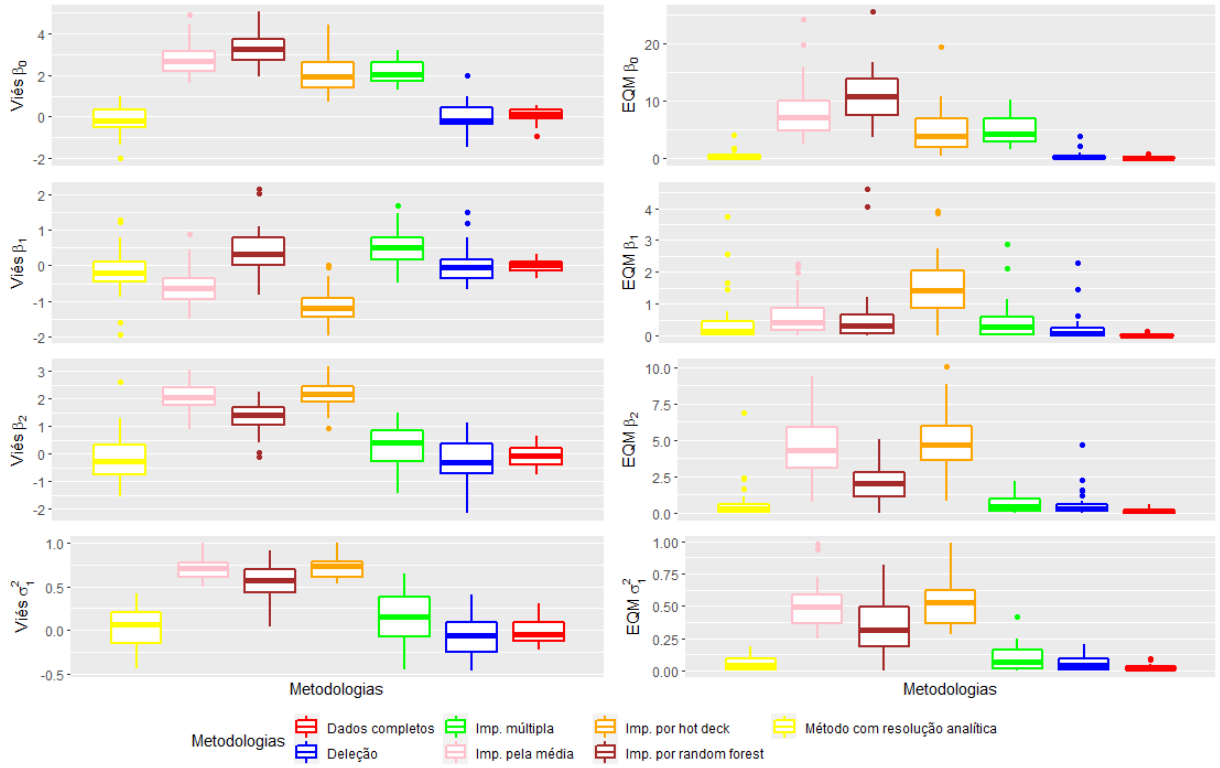

Figura 19 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário MNAR, $n = 100$ e $p = 0.60$.

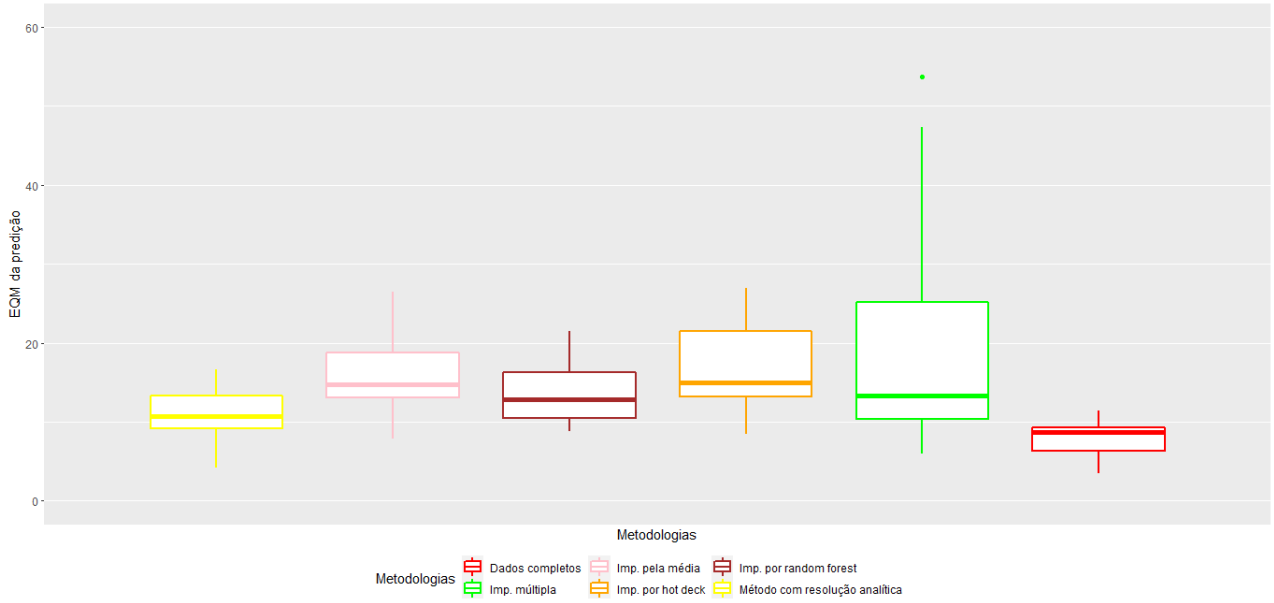

Figura 20 – Erro quadrático médio dos valores preditos para o cenário MNAR, $n = 100$ e $p = 0.60$.

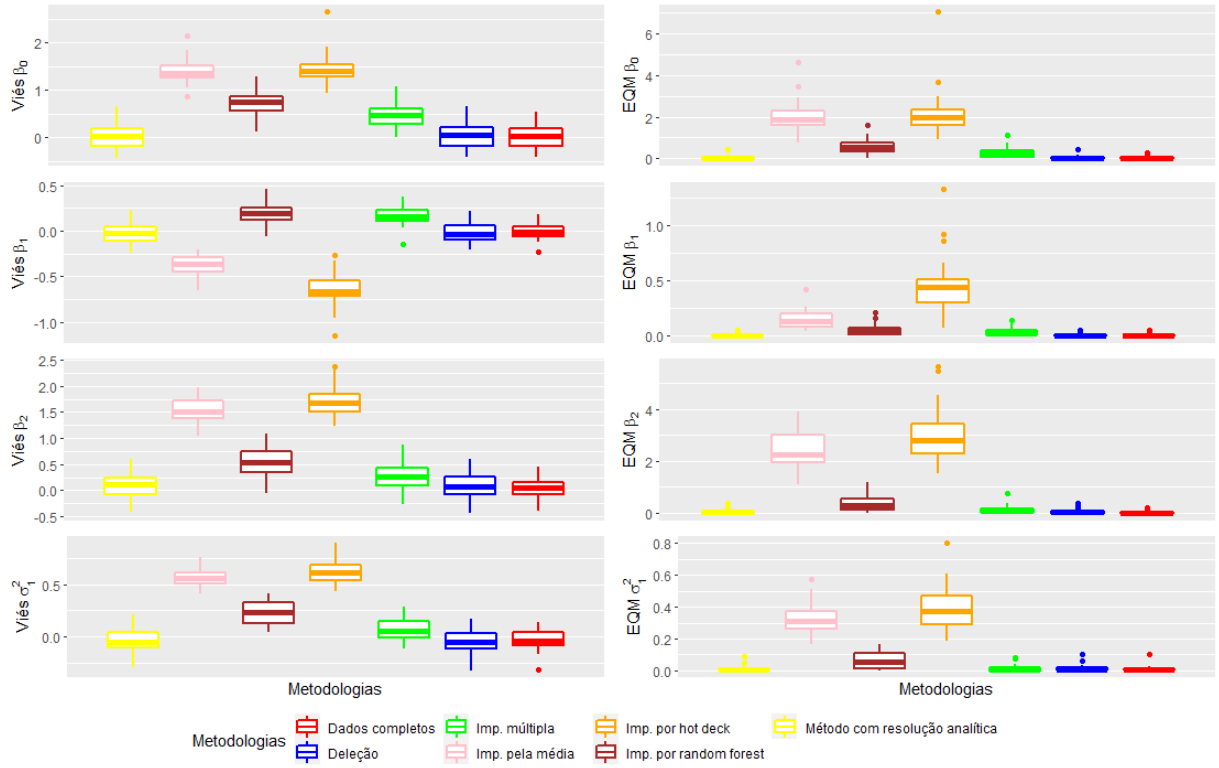

Figura 21 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário MNAR, $n = 300$ e $p = 0.20$.

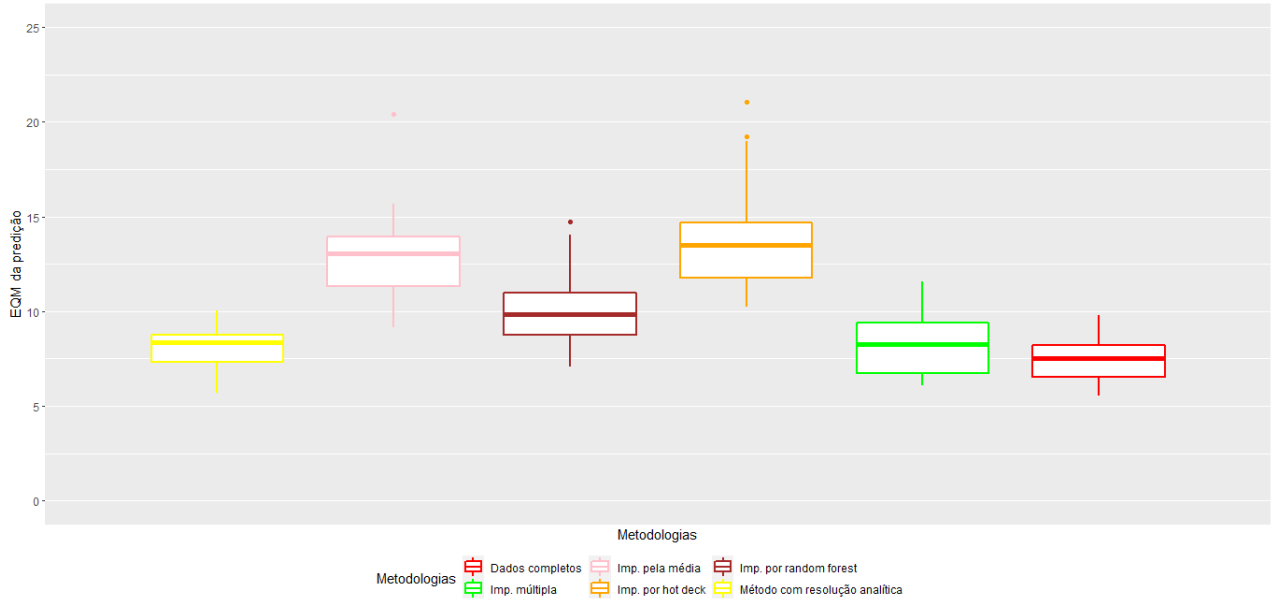

Figura 22 – Erro quadrático médio dos valores preditos para o cenário MNAR, $n = 300$ e $p = 0.20$.

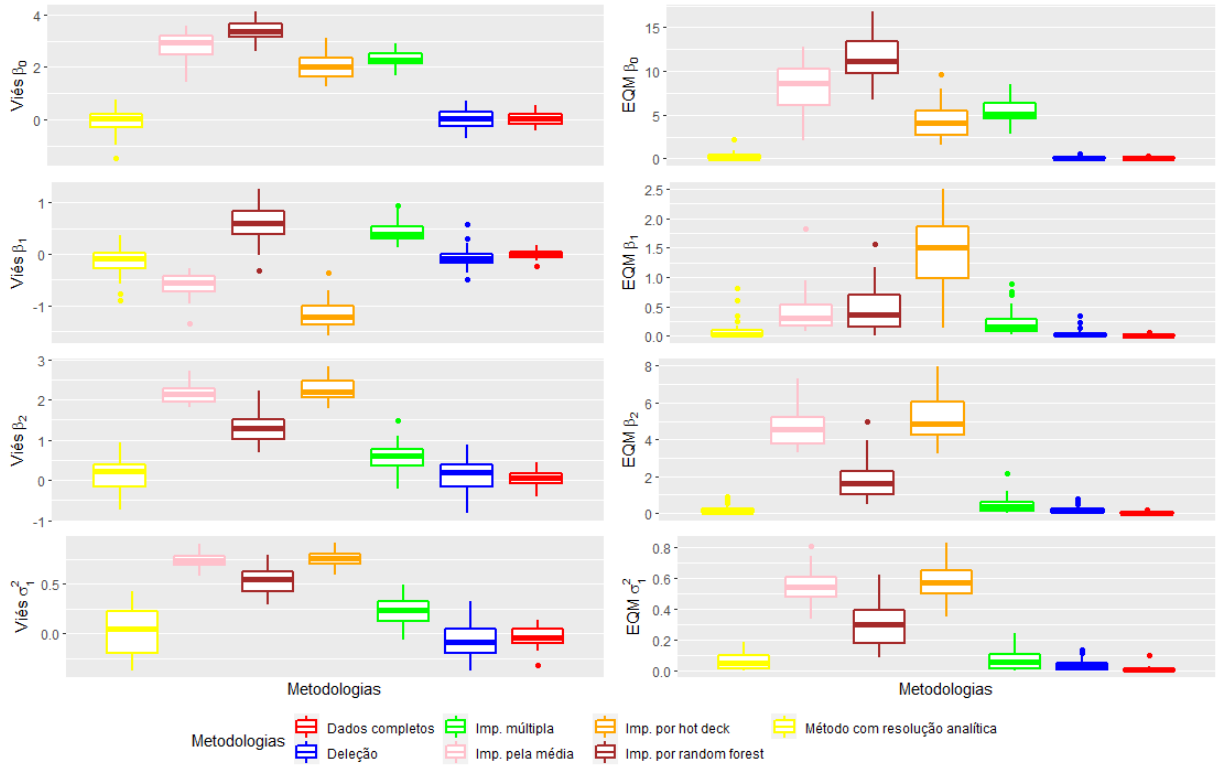

Figura 23 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário MNAR, $n = 300$ e $p = 0.60$.

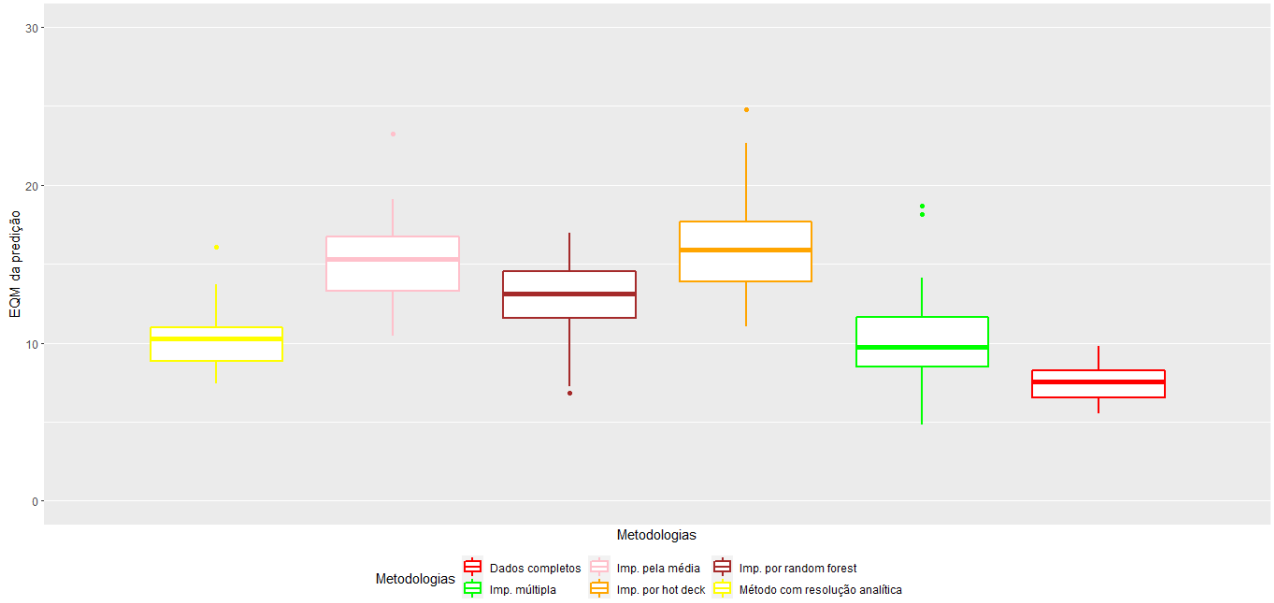

Figura 24 – Erro quadrático médio dos valores preditos para o cenário MNAR, $n = 300$ e $p = 0.60$.

Considerando o mecanismo MNAR de geração dos dados faltantes, que provavelmente é a situação de estimação e predição mais desafiadora, o método proposto apresenta um desempenho de estimação superior a todos os métodos comparados, exceto à estimação realizada com os dados completos, especialmente quando a proporção de dados faltantes é maior. O método de estimação baseado em modelos proposto nesse trabalho também apresenta os melhores desempenhos preditivos.

4.3 Modelo Gaussiano para duas variáveis com valores faltantes

Sejam Y, X_1, X_2, X_3 variáveis aleatórias, ou seja, $k = 3$. Vamos considerar que X_1 e X_2 possuem valores faltantes. Logo, como vimos na Seção 4.1, dada uma amostra de tamanho n de cada uma das variáveis Y, X_1, X_2, X_3, $(y_1, ..., y_n, x_{11}, ..., x_{1n}, x_{21}, ..., x_{2n}, x_{31}, ..., x_{3n})$ temos quatro cenários possíveis para esse caso: a) a observação i tem todas as suas variáveis observadas; b) a observação i tem valor faltante em X_1; c) a observação i tem valor faltante em X_2; d) a observação i tem valores faltantes em X_1 e X_2. Vamos analisar como ficam as funções densidade de probabilidade para cada um destes cenários, considerando o modelo Gaussiano:

a) A observação i tem todas as suas variáveis observadas.
Neste cenário, assumimos a distribuição de $Y|X_1, X_2, X_3$ de maneira similar à da Seção 4.2, com a diferença de que agora temos uma covariável a mais, X_3. Dessa forma, $f(y_i|x_{1i}, x_{2i}, x_{3i}, \theta)$ é a função densidade da distribuição $N\left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}, \sigma_1^2\right)$ e

$$f(y_i|x_{1i},x_{2i},x_{3i},\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{1}{2}\left(\frac{(y_i-(\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\beta_3 x_{3i}))^2}{\sigma_1^2}\right)\right); \tag{4.24}$$

b) A observação i tem valor faltante em X_1.
Neste cenário, como temos a obervação i faltante em X_1, queremos a distribuição de $Y|X_2, X_3$ que é obtida da mesma forma que a representada na Seção 4.2, apenas considerando uma covariável a mais, X_3. Logo:

$$f(y_i|x_{2i},x_{3i},\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi(\beta_1^2\sigma_2^2+\sigma_1^2)}} \exp\left(-\frac{1}{2}\left(\frac{(y_i-(\beta_0+\beta_2 x_{2i}+\beta_3 x_{3i}+(\gamma_0+\gamma_1 x_{2i}+\gamma_2 x_{3i})\beta_1)^2)}{\beta_1^2\sigma_2^2+\sigma_1^2}\right)\right); \tag{4.25}$$

c) A observação i tem valor faltante em X_2.
Neste cenário, como temos a obervação i faltante em X_2, queremos a distribuição de $Y|X_1, X_3$ que é obtida de maneira análoga a apresentada na Seção 4.2, apenas considerando uma covariável a mais, X_3 e, definindo $f(x_{2i}|x_{1i}, x_{3i}, \theta)$
como a função densidade da distribuição $N\left(\xi_0 + \xi_1 x_{1i} + \xi_2 x_{3i}, \sigma_3^2\right)$. Logo:

$$f(y_i|x_{1i},x_{3i},\boldsymbol{\theta})=\frac{1}{\sqrt{2\pi(\beta_2^2\sigma_3^2+\sigma_1^2)}}\exp\left(-\frac{1}{2}\left(\frac{(y_i-(\beta_0+\beta_1x_{1i}+\beta_3x_{3i}+(\xi_0+\xi_1x_{1i}+\xi_2x_{3i})\beta_2)^2)}{\beta_2^2\sigma_3^2+\sigma_1^2}\right)\right); \tag{4.26}$$

d) A observação i tem valores faltantes em X_1 e X_2.
Neste cenário, como temos a obervação i faltante em X_1 e X_2, queremos a distribuição de $Y|X_3$ que, de acordo com a Seção 4.1, é obtida por meio da integral dupla:

$$\begin{aligned} f(y_i|x_{3i},\boldsymbol{\theta}) &= \int\int f(y_i,x_{1i},x_{2i}|x_{3i},\boldsymbol{\theta})dx_{1i}dx_{2i} \\ &= \int\int f(y_i|x_{1i},x_{2i},x_{3i},\boldsymbol{\theta})f(x_{1i}|x_{2i},x_{3i},\boldsymbol{\theta})f(x_{2i}|x_{3i},\boldsymbol{\theta})dx_{1i}dx_{2i} \\ &= \int\left[\int f(y_i|x_{1i},x_{2i},x_{3i},\boldsymbol{\theta})f(x_{1i}|x_{2i},x_{3i},\boldsymbol{\theta})dx_{1i}\right]f(x_{2i}|x_{3i},\boldsymbol{\theta})dx_{2i}. \end{aligned} \tag{4.27}$$

Sabemos que a primeira integral em relação a x_{1i} da Equação (4.27) se refere à distribuição de $Y|X_2,X_3$. Logo, substituindo esta integral pela expressão da $f(y_i|x_{2i},x_{3i},\boldsymbol{\theta})$ denotada no item b), temos:

$$\begin{aligned} f(y_i|x_{3i},\boldsymbol{\theta}) = \int & \left[\frac{1}{\sqrt{2\pi(\beta_1^2\sigma_2^2+\sigma_1^2)}}\right. \\ & \times \exp\left(-\frac{1}{2}\left(\frac{(y_i-(\beta_0+\beta_2x_{2i}+\beta_3x_{3i}+(\gamma_0+\gamma_1x_{2i}+\gamma_2x_{3i})\beta_1)^2)}{\beta_1^2\sigma_2^2+\sigma_1^2}\right)\right) \\ & \times \left. f(x_{2i}|x_{3i},\boldsymbol{\theta})\right]dx_{2i}. \end{aligned} \tag{4.28}$$

Assumindo $f(x_{2i}|x_{3i},\boldsymbol{\theta})$ como a função densidade da distribuição $N(\mu_0+\mu_1x_{3i},\sigma_4^2)$, temos:

$$
\begin{aligned}
f(y_i|x_{3i},\theta) &= \int \left[\frac{1}{\sqrt{2\pi(\beta_1^2\sigma_2^2+\sigma_1^2)}} \right. \\
&\times \exp\left(-\frac{1}{2}\left(\frac{(y_i-(\beta_0+\beta_2 x_{2i}+\beta_3 x_{3i}+(\gamma_0+\gamma_1 x_{2i}+\gamma_2 x_{3i})\beta_1)^2)}{\beta_1^2\sigma_2^2+\sigma_1^2}\right)\right) \\
&\times \left. \frac{1}{\sqrt{2\pi\sigma_4^2}} \exp\left(-\frac{1}{2}\left(\frac{(x_{2i}-(\mu_0+\mu_1 x_{3i}))^2}{\sigma_4^2}\right)\right)\right] dx_{2i} \\
&= \int \left[\frac{1}{\sqrt{2\pi(\beta_1^2\sigma_2^2+\sigma_1^2)}} \right. \\
&\times \exp\left(-\frac{1}{2}\left(\frac{(y_i-\beta_0-\beta_3 x_{3i}-\gamma_0\beta_1-\gamma_2\beta_1 x_{3i}-(\beta_2+\gamma_1\beta_1)x_{2i})^2}{\beta_1^2\sigma_2^2+\sigma_1^2}\right)\right) \\
&\times \left. \frac{1}{\sqrt{2\pi\sigma_4^2}} \exp\left(-\frac{1}{2}\left(\frac{(x_{2i}-(\mu_0+\mu_1 x_{3i}))^2}{\sigma_4^2}\right)\right)\right] dx_{2i}
\end{aligned}
$$

e fazendo $A = y_i-\beta_0-\beta_3 x_{3i}-\gamma_0\beta_1-\gamma_2\beta_1 x_{3i},\ B=\beta_2+\gamma_1\beta_1,\ C=\beta_1^2\sigma_2^2+\sigma_1^2$ e $D=\mu_0+\mu_1 x_{3i}$, temos:

$$
\begin{aligned}
f(y_i|x_{3i},\theta) &= \frac{1}{2\pi\sqrt{C\sigma_4^2}} \int \exp\left(-\tfrac{1}{2}\left(\tfrac{(A-Bx_{2i})^2}{C}\right)\right) \exp\left(-\tfrac{1}{2}\left(\tfrac{(x_{2i}-D)^2}{\sigma_4^2}\right)\right) dx_{2i} \\
&= \frac{1}{2\pi\sqrt{C\sigma_4^2}} \int \exp\left(\frac{-(B^2\sigma_4^2+C)x_{2i}^2+2(AB\sigma_4^2+DC)x_{2i}-(A^2\sigma_4^2+D^2C)}{2C\sigma_4^2}\right) dx_{2i}.
\end{aligned}
$$

Assumindo $F=B^2\sigma_4^2+C$, $G=AB\sigma_4^2+DC$ e $H=A^2\sigma_4^2+D^2C$, temos:

$$
\begin{aligned}
f(y_i|x_{3i},\theta) &= \frac{1}{2\pi\sqrt{C\sigma_4^2}} \exp\left(-\tfrac{H}{2C\sigma_4^2}\right) \int \exp\left(\tfrac{-Fx_2^2+2Gx_2}{2C\sigma_4^2}\right) dx_{2i} \\
&= \frac{1}{2\pi\sqrt{C\sigma_4^2}} \exp\left(-\tfrac{H}{2C\sigma_4^2}\right) \exp\left(\tfrac{G^2}{2CF\sigma_4^2}\right) \int \exp\left(-\tfrac{1}{2}\left(\tfrac{(x_{2i}-\frac{G}{F})^2}{\frac{C}{F}\sigma_4^2}\right)\right) dx_{2i} \\
&= \frac{1}{2\pi\sqrt{C\sigma_4^2}} \exp\left(-\tfrac{H}{2C\sigma_4^2}\right) \exp\left(\tfrac{G^2}{2CF\sigma_4^2}\right) \sqrt{\tfrac{2\pi C\sigma_4^2}{F}} \int \tfrac{1}{\sqrt{\frac{2\pi C\sigma_4^2}{F}}} \exp\left(-\tfrac{1}{2}\left(\tfrac{(x_{2i}-\frac{G}{F})^2}{\frac{C}{F}\sigma_4^2}\right)\right) dx_{2i} \\
&= \frac{1}{2\pi\sqrt{C\sigma_4^2}} \exp\left(-\tfrac{H}{2C\sigma_4^2}\right) \exp\left(\tfrac{G^2}{2CF\sigma_4^2}\right) \sqrt{\tfrac{2\pi C\sigma_4^2}{F}} \\
&= \frac{1}{\sqrt{2\pi F}} \exp\left(\tfrac{-HF+G^2}{2CF\sigma_4^2}\right) \\
&= \frac{1}{\sqrt{2\pi F}} \exp\left(-\tfrac{1}{2}\left(\tfrac{(y_i-(\beta_0+\beta_3 x_{3i}+\gamma_0\beta_1+\gamma_2\beta_1 x_{3i}+\beta_2\mu_0+\beta_2\mu_1 x_{3i}+\gamma_1\beta_1\mu_0+\gamma_1\beta_1\mu_1 x_{3i}))^2}{F}\right)\right).
\end{aligned}
$$

Portanto, $f(y_i|x_{3i},\theta)$ tem a seguinte expressão:

$$\frac{1}{\sqrt{2\pi F}}\exp\left(-\frac{1}{2}\left(\frac{(y_i-(\beta_0+\beta_3 x_{3i}+\gamma_0\beta_1+\gamma_2\beta_1 x_{3i}+\beta_2\mu_0+\beta_2\mu_1 x_{3i}+\gamma_1\beta_1\mu_0+\gamma_1\beta_1\mu_1 x_{3i}))^2}{F}\right)\right)$$

em que, $F = B^2\sigma_4^2 + C = \beta_2^2\sigma_4^2 + 2\beta_2\beta_1\gamma_1\sigma_4^2 + \gamma_1^2\beta_1^2\sigma_4^2 + \beta_1^2\sigma_2^2 + \sigma_1^2$.

Vamos construir a função de verossimilhança deste modelo, com base no que foi discutido na Seção 4.1, da seguinte forma:

$$\begin{aligned} L(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) = \prod_{i=1}^{n} & \Big[f(y_i|x_{1i},x_{2i},x_{3i},\theta)^{\delta_1(i)\delta_2(i)} \times f(y_i|x_{2i},x_{3i},\theta)^{(1-\delta_1(i))\delta_2(i)} \\ & \times \quad f(y_i|x_{1i},x_{3i},\theta)^{(1-\delta_2(i))\delta_1(i)} \times f(y_i|x_{3i},\theta)^{(1-\delta_1(i))(1-\delta_2(i))}\Big] \end{aligned} \tag{4.29}$$

Observe que o espaço paramétrico θ da função de verossimilhança expressa na Equação 4.29 para este caso em que temos uma regressão linear múltipla com três variáveis explicativas, sendo duas delas com valores faltantes, se torna
$\theta = (\beta_0,\beta_1,\beta_2,\beta_3,\gamma_0,\gamma_1,\gamma_2,\xi_0,\xi_1,\xi_2,\mu_0,\mu_1,\sigma_1^2,\sigma_2^2,\sigma_3^2,\sigma_4^2)$.

Para obtermos a função de log-verossimilhança $l(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3)$, basta fazermos $\log L(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3)$:

$$\begin{aligned} l(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) &= \log L(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) \\ &= \log \Bigg[\prod_{i=1}^{n}\Big[f(y_i|x_{1i},x_{2i},x_{3i},\theta)^{\delta_1(i)\delta_2(i)} \times f(y_i|x_{2i},x_{3i},\theta)^{(1-\delta_1(i))\delta_2(i)} \\ &\times \quad f(y_i|x_{1i},x_{3i},\theta)^{(1-\delta_2(i))\delta_1(i)} \times f(y_i|x_{3i},\theta)^{(1-\delta_1(i))(1-\delta_2(i))}\Big]\Bigg] \\ &= \sum_{i=1}^{n} \quad \delta_1(i)\delta_2(i)\log f(y_i|x_{1i},x_{2i},x_{3i},\theta) + (1-\delta_1(i))\delta_2(i)\log f(y_i|x_{2i},x_{3i},\theta) \\ &+ \quad (1-\delta_2(i))\delta_1(i)\log f(y_i|x_{1i},x_{3i},\theta) + (1-\delta_1(i))(1-\delta_2(i))\log f(y_i|x_{3i},\theta). \end{aligned} \tag{4.30}$$

Para encontrarmos as estimativas dos parâmetros, temos que maximizar a função de log-verossimilhança dada pela Equação (4.30), em que as densidades $f(y_i|x_{1i},x_{2i},x_{3i},\theta)$, $f(y_i|x_{2i},x_{3i},\theta)$, $f(y_i|x_{1i},x_{3i},\theta)$ e $f(y_i|x_{3i},\theta)$ são denotadas nos itens a), b), c) e d) desta seção.

4.3.1 Análise preditiva do método

Para analisarmos a capacidade preditiva do método proposto, dividimos o banco de dados em 70% para treino (subconjunto para o qual estimamos os parâmetros) e 30% para teste (subconjunto através do qual calculamos os valores preditos $\widehat{y}$ e comparamos com os observados y). Para o cálculo do $\widehat{y}$ consideramos o valor esperado estimado da distribuição de $Y|X_1,X_2,X_3$, de $Y|X_2,X_3$, de $Y|X_1,X_3$ ou de $Y|X_3$, de acordo com qual cenário a), b), c) ou d), descritos na Seção 4.3, uma observação i pertença. Logo,

i) Se a informação da observação i é completa em relação a X_1 e X_2, temos:

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_{1i} + \widehat{\beta_2} x_{2i} + \widehat{\beta_3} x_{3i}, \tag{4.31}$$

em que $\widehat{\beta_0}$, $\widehat{\beta_1}$, $\widehat{\beta_2}$ e $\widehat{\beta_3}$ são os valores das estimativas dos parâmetros e, como $f(y_i|x_{1i},x_{2i},x_{3i},\theta)$ é a função densidade da distribuição $N\left(\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\beta_3 x_{3i},\sigma_1^2\right)$, então $\widehat{y_i}$ dado pela Equação (4.31) se refere ao valor esperado estimado da distribuição de $Y|X_1,X_2,X_3$;

ii) Se a informação da observação i é faltante apenas em relação a X_1, temos:

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_2} x_{2i} + \widehat{\beta_3} x_{3i} + (\widehat{\gamma_0} + \widehat{\gamma_1} x_{2i} + \widehat{\gamma_2} x_{3i})\widehat{\beta_1}, \tag{4.32}$$

em que $\widehat{\beta_0}$, $\widehat{\beta_1},\widehat{\beta_2}$, $\widehat{\beta_3},\widehat{\gamma_0}$, $\widehat{\gamma_1}$ e $\widehat{\gamma_2}$ são os valores das estimativas dos parâmetros e, como $f(y_i|x_{2i},x_{3i},\theta)$ é a função densidade da distribuição $N\left(\beta_0+\beta_2 x_{2i}+\beta_3 x_{3i}+(\gamma_0+\gamma_1 x_{2i}+\gamma_2 x_{3i})\beta_1,\beta_1^2\sigma_2^2+\sigma_1^2\right)$, então $\widehat{y_i}$ dado pela Equação (4.32) se refere ao valor esperado estimado da distribuição de $Y|X_2,X_3$;

iii) Se a informação da observação i é faltante apenas em relação a X_2, temos:

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_{1i} + \widehat{\beta_3} x_{3i} + (\widehat{\xi_0} + \widehat{\xi_1} x_{1i} + \widehat{\xi_2} x_{3i})\widehat{\beta_2}, \tag{4.33}$$

em que $\widehat{\beta_0}$, $\widehat{\beta_1},\widehat{\beta_2}$, $\widehat{\beta_3},\widehat{\xi_0}$, $\widehat{\xi_1}$ e $\widehat{\xi_2}$ são os valores das estimativas dos parâmetros e, como $f(y_i|x_{1i},x_{3i},\theta)$ é a função densidade da distribuição $N\left(\beta_0+\beta_1 x_{1i}+\beta_3 x_{3i}+(\xi_0+\xi_1 x_{1i}+\xi_2 x_{3i})\beta_2,\beta_2^2\sigma_3^2+\sigma_1^2\right)$, então $\widehat{y_i}$ dado pela Equação (4.33) se refere ao valor esperado estimado da distribuição de $Y|X_1,X_3$;

iv) Se a informação da observação i é faltante em relação a X_1 e X_2, temos:

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_3} x_{3i} + \widehat{\gamma_0}\widehat{\beta_1} + \widehat{\gamma_2}\widehat{\beta_1} x_{3i} + \widehat{\beta_2}\widehat{\mu_0} + \widehat{\beta_2}\widehat{\mu_1} x_{3i} + \widehat{\gamma_1}\widehat{\beta_1}\widehat{\mu_0} + \widehat{\gamma_1}\widehat{\beta_1}\widehat{\mu_1} x_{3i}, \tag{4.34}$$

em que $\widehat{\beta_0}$, $\widehat{\beta_1},\widehat{\beta_2}$, $\widehat{\beta_3},\widehat{\gamma_0}$, $\widehat{\gamma_1}$, $\widehat{\mu_0}$ e $\widehat{\mu_1}$ são os valores das estimativas dos parâmetros e, como $f(y_i|x_{3i},\theta)$ é a função densidade da distribuição $N(\beta_0+\beta_3 x_{3i}+\gamma_0\beta_1+\gamma_2\beta_1 x_{3i}+\beta_2\mu_0+\beta_2\mu_1 x_{3i}+\gamma_1\beta_1\mu_0+\gamma_1\beta_1\mu_1 x_{3i},F)$, em que $F = B^2\sigma_4^2+C=\beta_2^2\sigma_4^2+2\beta_2\beta_1\gamma_1\sigma_4^2+\gamma_1^2\beta_1^2\sigma_4^2+\beta_1^2\sigma_2^2+\sigma_1^2$, então $\widehat{y_i}$ dado pela Equação (4.34) se refere ao valor esperado estimado da distribuição de $Y|X_3$.

Efetuado o cálculo do $\widehat{y}$ para as observações da base de teste, calculamos a média das diferenças quadráticas entre $\widehat{y}$ e y, o erro quadrático médio da predição. Quanto mais próximo de zero estiver essa soma quadrática, melhor é a predição do modelo estimado.

4.3.2 Estudo de simulação

Nesta seção, será discutido como foi feito o estudo de simulação e comparação do desempenho do método proposto com o de outros métodos de imputação e deleção de dados.

Nesse cenário com duas variáveis com valores faltantes, também comparamos o desempenho dos diferentes métodos em relação ao viés e ao erro quadrático médio (EQM) dos valores das estimativas dos parâmetros e em relação ao poder preditivo definido pela média das diferenças quadráticas entre $\widehat{y}$ e y em amostras teste, separadas especificamente para esse fim.

Vários cenários de simulação foram testados, entre os quais variamos: o tamanho da amostra ($n = 100$ e $n = 300$), a proporção de valores faltantes tanto em relação a X_1 quanto em relação a X_2 ($p = 0.20$ e $p = 0.60$) e o mecanismo que gera os dados faltantes, podendo ser MCAR para X_1 e X_2, MAR para X_1 e X_2 ou MNAR para X_1 e X_2. Para cada cenário analisado, simulamos 30 réplicas (amostras diferentes) sob as mesmas condições. Com elas, temos amostras de tamanho 30 para conduzir análises de desempenho através do viés e EQM das estimativas dos parâmetros, assim como do erro de predição. Quanto mais próximas de zero essas diferenças, mais precisa é a estimação e a predição. Aqui, adotamos o modelo normal como a distribuição de probabilidades das variáveis, mas qualquer outro modelo que possibilite a resolução analítica pode ser considerado e a metodologia adaptada.

Após a geração do conjunto de dados, separamos as 70% primeiras observações para treino e estimação, através da qual fazemos a análise inferencial dos parâmetros e os outros 30% para teste, em que analisamos o método de estimação baseado em modelo na presença de valores faltantes em relação ao poder preditivo, comparando-o com os métodos: modelo completo (sem dados faltantes), método de deleção *listwise*, método de imputação pela média, método de imputação por *Random Forest*, método de imputação por *hot deck* e método de imputação múltipla.

Realizamos as simulações de acordo com os seguintes passos:

Passo 1: Geramos x_{3i} da distribuição Normal com média 0 e variância 1; geramos x_{2i} da distribuição Normal de média $\mu_0 + \mu_1 x_{3i}$ e variância σ_4^2; geramos x_{1i} da distribuição Normal de média $\gamma_0 + \gamma_1 x_{2i} + \gamma_2 x_{3i}$ e variância σ_2^2 e geramos y_i da distribuição Normal de média $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}$ e variância σ_1^2, com $i = 1, ..., n$, sendo n, portanto, o tamanho da amostra final que iremos obter. O verdadeiros valores destes parâmetros considerados são: $\sigma_1^2 = 4,\ \sigma_2^2 = 8,\ \sigma_3^2 = 8,\ \sigma_4^2 = 8,\ \beta_0 = -1.1,\ \beta_1 = 1.8,\ \beta_2 = 1.1,\ \beta_3 = 1.5,\ \gamma_0 = 1.8,\ \gamma_1 = -1.1,\ \gamma_2 = -1.1,\ \mu_0 = -1.1$ e $\mu_1 = 1.2$;

Passo 2: De acordo com o mecanismo considerado, geramos valores faltantes nas variáveis X_1 e X_2 da seguinte forma:

a) Se o mecanismo é MCAR: geramos n valores da distribuição Uniforme entre 0 e 1 e, se o valor na posição $j,\ j \leq n$, for menor ou igual a porcentagem $0 \leq p \leq 1$ que queremos obter de valores faltantes, então, o j-ésimo valor da variável X_1 receberá NA. Análogo para a geração de valores faltantes da variável X_2;

b) Se o mecanismo é MAR: Escolhemos os $p \times n$ maiores valores da variável X_3 e, para as observações X_1 correspondentes a estes valores, retornamos NA. Análogo para a geração de valores faltantes da variável X_2;

c) Se o mecanismo é MNAR: Escolhemos os $p \times n$ maiores valores da variável X_1 e, para estes valores, retornamos NA. Para a geração de valores faltantes da variável X_2, escolhemos os $p \times n$ maiores valores da variável X_2 e, para estes valores, retornamos NA.

Ressaltamos que no caso de estimação pelo modelo completo, para comparação de desempenho, esse passo não é realizado;

Passo 3: Maximizamos a função de log-verossimilhança da mesma forma que descrito na Seção 4.2.2. Quanto às funções de log-verossimilhança maximizadas, temos que:

a) Para o método baseado em modelo aqui proposto, a função de log-verossimilhança será a construída na Seção 4.3, em que consideramos os quatro cenários possíveis para cada observação do conjunto de treinamento, ou seja, a observação i ser observada em relação à variável X_1 e ser faltante em relação à variável X_2; a observação i ser observada em relação à variável X_2 e ser faltante em relação à variável X_1; a observação i ser observada em relação às variáveis X_1 e X_2 ou a observação i ser faltante em relação às variáveis X_1 e X_2;

b) Para o método de imputação pela média, primeiramente calculamos a média dos valores do conjunto de treinamento correspondentes às variáveis X_1 e X_2 que não estão faltantes e, em seguida, para as observações que possuem valores faltantes em X_1 e X_2, imputamos as respectivas médias. Depois, com o conjunto de dados completo obtido com estas imputações, maximizamos a função de log-verossimilhança da Seção 4.3 referente ao cenário em que as observações são completas em relação a X_1 e X_2;

c) Para o método de imputação por *Random Forest*, imputamos os valores faltantes em relação a X_1 e X_2 utilizando o pacote *missForest*. Após, com o conjunto de dados completo obtido com esta imputação, maximizamos a função de log-verossimilhança da Seção 4.3 referente ao cenário em que as observações são completas em relação a X_1 e X_2;

d) Para o método de imputação por *Hot-Deck*, imputamos os valores faltantes em relação a X_1 e X_2 utilizando o pacote *VIM*. Após, com o conjunto de dados completo, obtido com esta imputação, maximizamos a função de log-verossimilhança da Seção 4.3 referente ao cenário em que as observações são completas em relação a X_1 e X_2;

e) Para o método de imputação múltipla, imputamos os valores faltantes em relação a X_1 e X_2 utilizando o pacote *Amelia*. Por meio deste pacote, criamos cinco conjuntos de dados completos e, para encontrarmos a estimativa final dos parâmetros, calculamos

a média aritmética das cinco estimativas para cada conjunto de dados completos (uma para cada conjunto de dados completos) encontradas por meio da maximização da função de log-verossimilhança da Seção 4.3 referente ao cenário em que as observações são completas em relação a X_1 e X_2;

f) Para o método de deleção de casos, foram removidas todas as observações que tinham valores faltantes em X_1 ou X_2 e então, com a subamostra de valores completos restante, maximizamos a função de log-verossimilhança da Seção 4.3 referente a este cenário;

g) Por fim, consideramos o caso em que não teríamos valores faltantes, utilizamos o conjunto de dados completo para maximizarmos a função de log-verossimilhança da Seção 4.3 referente a este cenário.

Passo 4: Após obtermos os valores das estimativas dos parâmetros para cada conjunto dentro de cada método, calculamos a diferença e a diferença quadrática dessas estimativas em relação aos verdadeiros valores dos parâmetros. Como os parâmetros σ_2^2, σ_3^2, σ_4^2, γ_0, γ_1, γ_2, μ_0 e μ_1 são estimados diretamente apenas pela metodologia proposta, os índices de desempenho dos seus estimadores não são mostrados e comparados;

Passo 5: Para analisarmos os métodos em relação ao poder preditivo, calculamos o erro quadrático médio do $\widehat{y}_i$ em relação ao y_i observado para todas as observações do conjunto de teste. O cálculo de $\widehat{y}_i$ se dá da seguinte forma:

a) Para a metodologia proposta e cada observação i do conjunto de dados de teste, calculamos $\widehat{y}_i$ de acordo com o proposto na Seção 4.3.1;

b) Para os métodos de imputação de dados por *Random Forest*, imputação por *Hot-Deck* e imputação múltipla, realizamos a imputação dos dados faltantes na base teste usando os mesmos procedimentos do passo 3 e, com os dados completos, calculamos $\widehat{y}_i$ de acordo com o item i) da Sessão 4.3.1. Para o caso da imputação múltipla, como criamos cinco conjuntos de dados completos, o erro quadrático médio é dado pela média entre os erros quadráticos médios dos cinco conjuntos gerados;

c) Para o método de imputação pela média, a média das observações não faltantes em X_1 do conjunto de treino é o valor utilizado para ser imputado nas observações que possuem valores faltantes no conjunto de teste para a variável X_1. Analogamente, a média das observações não faltantes em X_2 do conjunto de treino é o valor utilizado para ser imputado nas observações que possuem valores faltantes no conjunto de teste para a variável X_2. Após esta imputações, calculamos $\widehat{y}_i$ de acordo com o item i) da Sessão 4.3.1;

d) Para o método em que consideramos o conjunto de dados de teste completo, sem valores faltantes, calculamos $\widehat{y}_i$ de acordo com o item i) da Sessão 4.3.1;

e) Para o método de deleção de dados, como deletamos as observações que possuem valores faltantes, não conseguimos calcular o $\widehat{y}_i$ para eles pois não existe, de fato, um processo de imputação ou predição de valores faltantes. Logo, não analisamos o desempenho de predição desse método nas observações do conjunto de teste, por não fazer sentido essa comparação.

Seguem abaixo os resultados dos cenários de simulação. As Figuras 25, 27, 29 e 31 apresentam o desempenho inferência dos métodos para o mecanismo MCAR de dados faltantes e as Figuras 26, 28, 30 e 32 o desempenho preditivo.

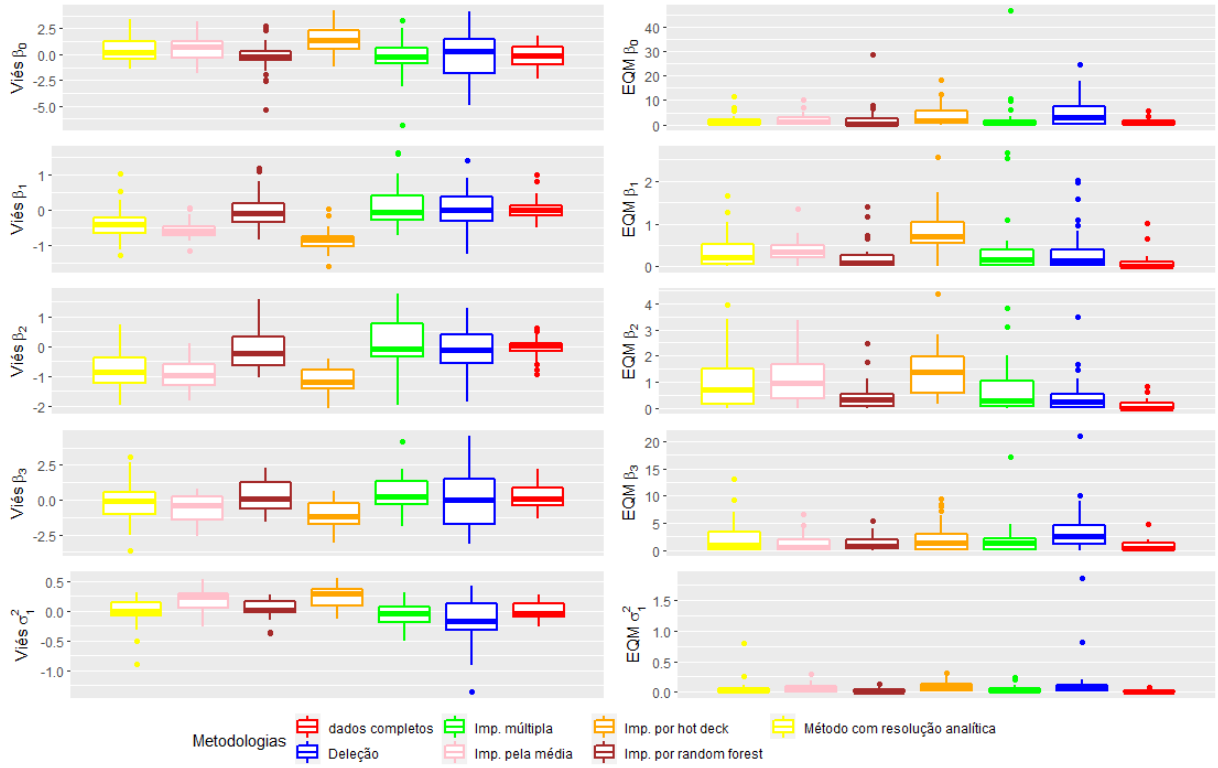

Figura 25 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário com X_1 e X_2 MCAR, $n = 100$, $p = 0.20$ para X_1 e $p = 0.60$ para X_2.

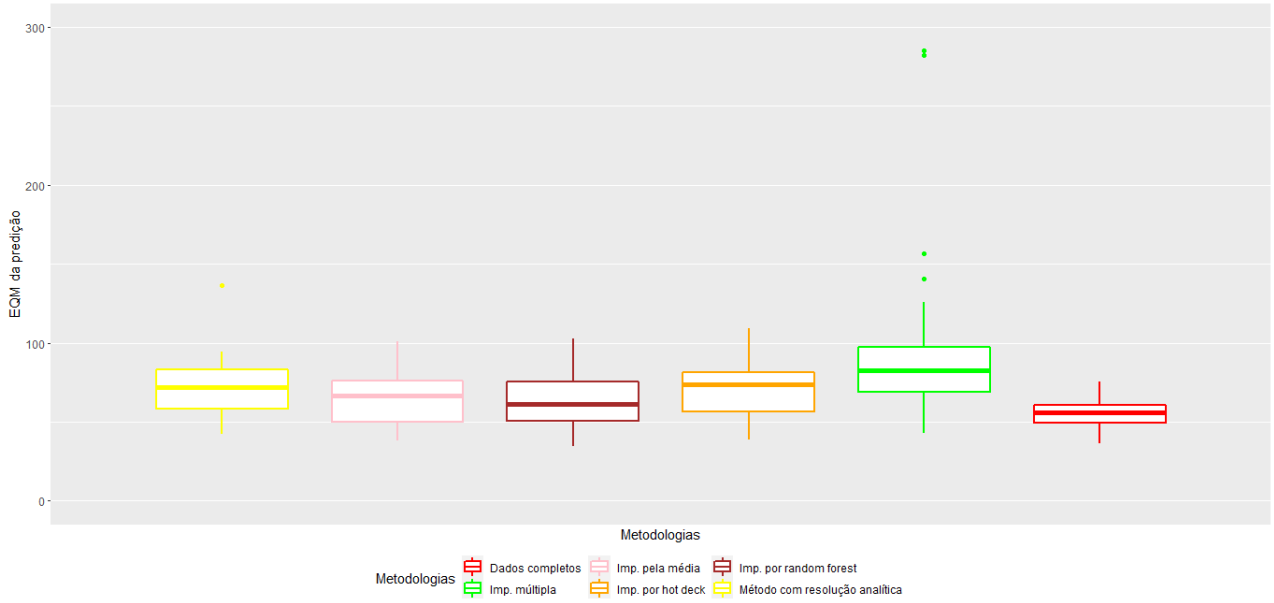

Figura 26 – Erro quadrático médio dos valores preditos para o cenário com X_1 e X_2 MCAR, $n = 100$, $p = 0.20$ para X_1 e $p = 0.60$ para X_2.

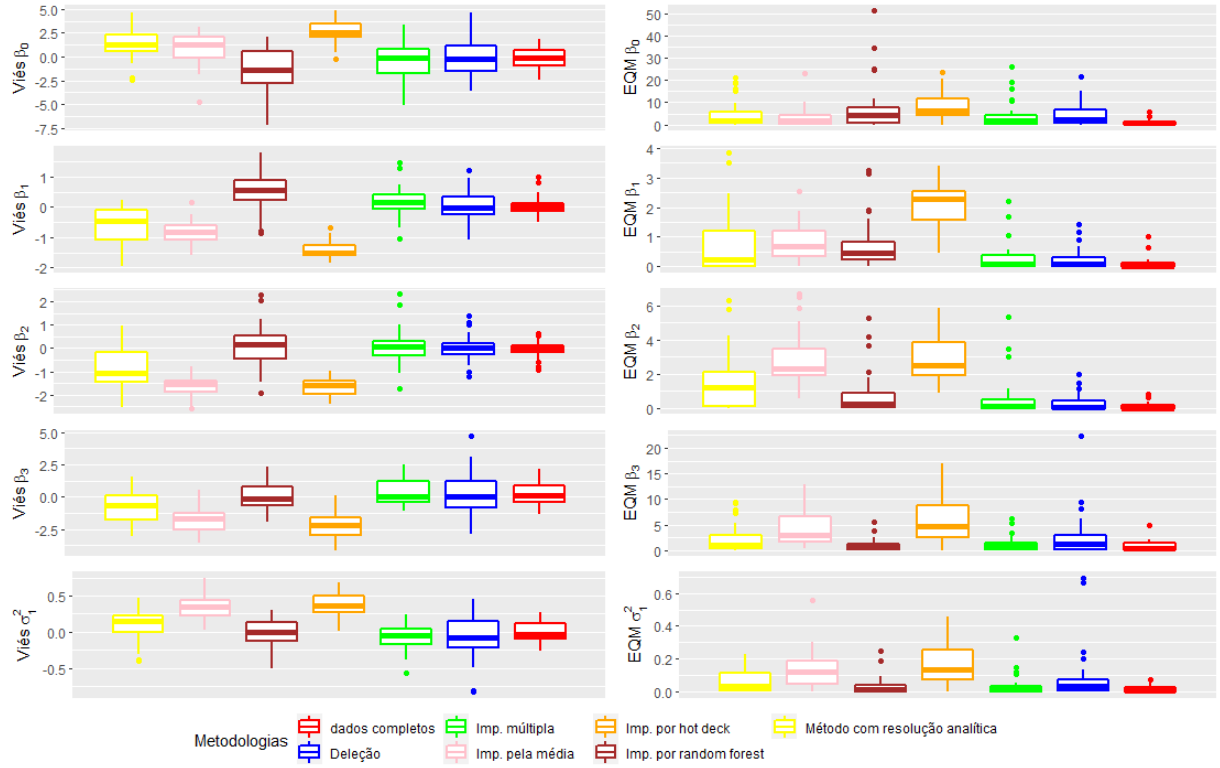

Figura 27 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário com X_1 e X_2 MCAR, $n = 100$, $p = 0.60$ para X_1 e $p = 0.20$ para X_2.

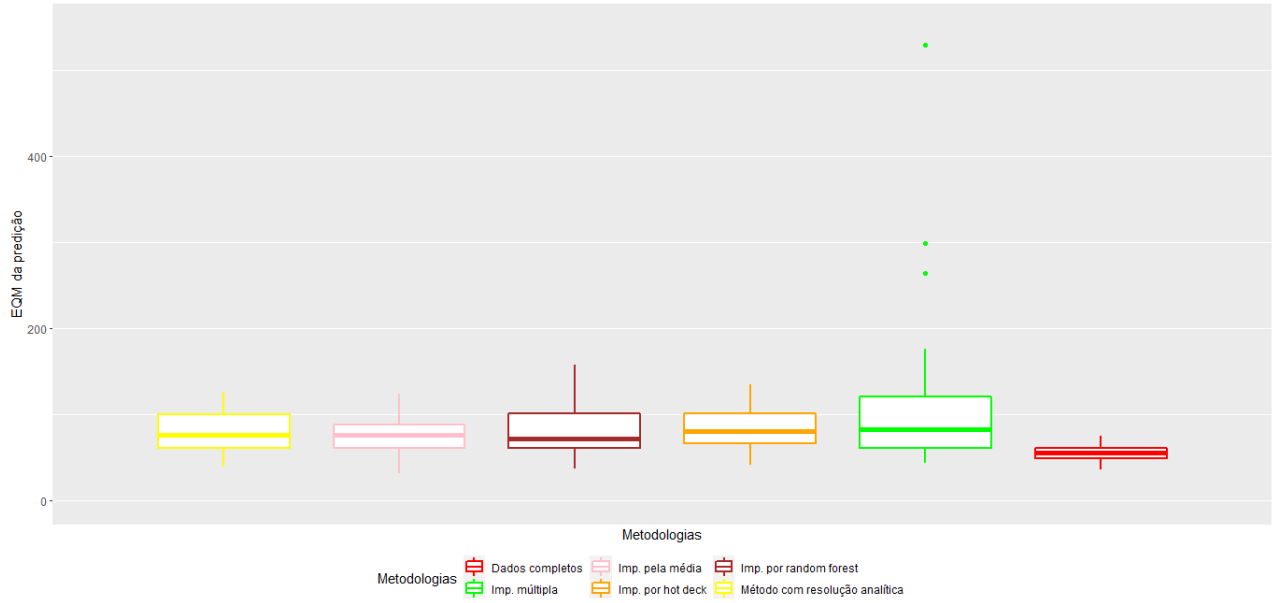

Figura 28 – Erro quadrático médio dos valores preditos para o cenário com X_1 e X_2 MCAR, $n = 100$, $p = 0.60$ para X_1 e $p = 0.20$ para X_2.

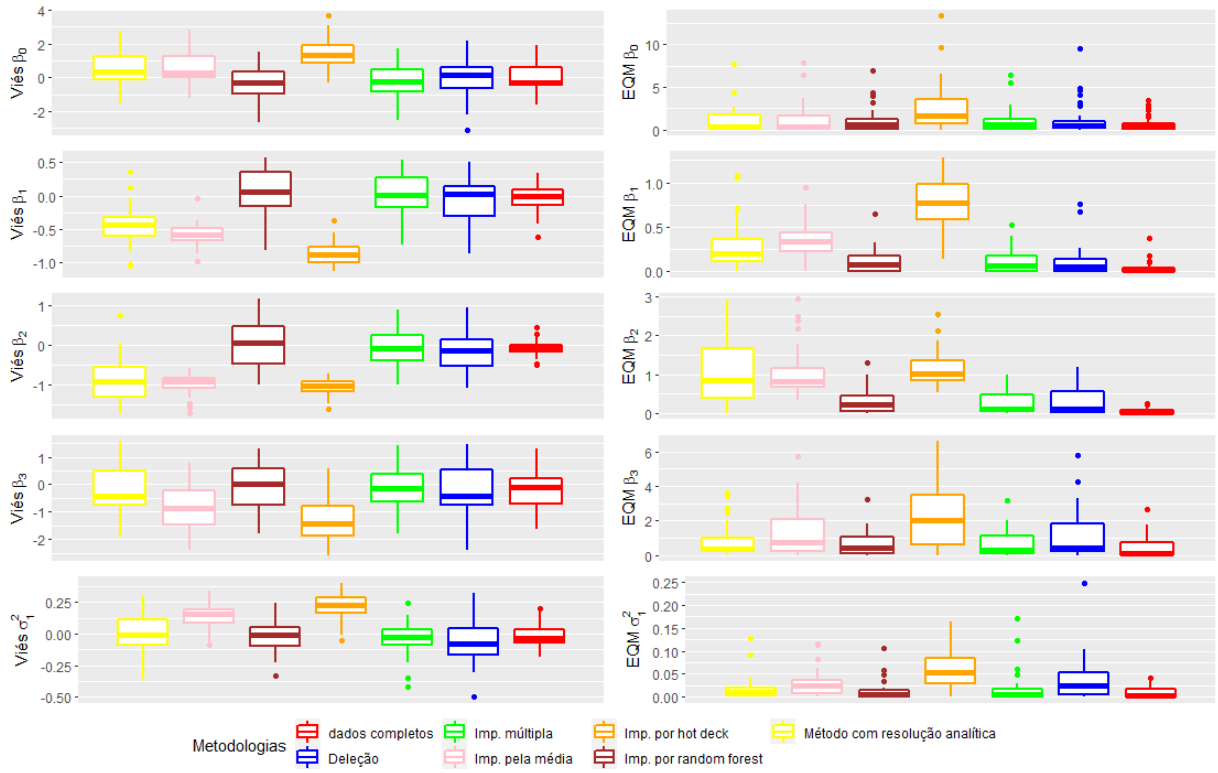

Figura 29 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário com X_1 e X_2 MCAR, $n = 300$, $p = 0.20$ para X_1 e $p = 0.60$ para X_2.

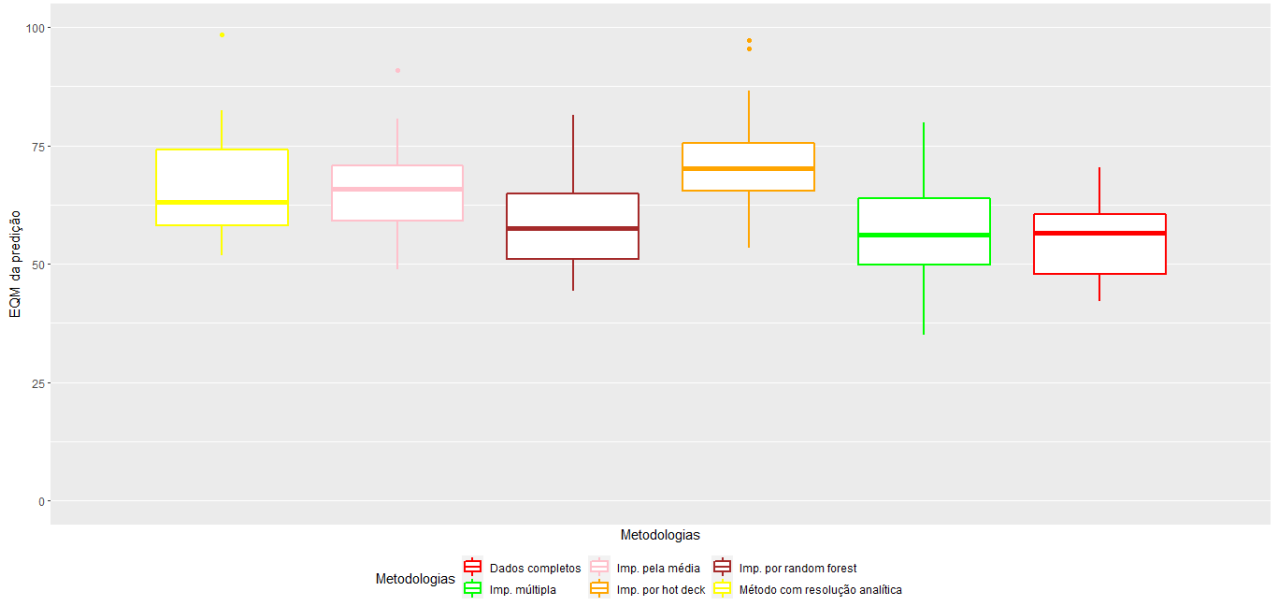

Figura 30 – Erro quadrático médio dos valores preditos para o cenário com X_1 e X_2 MCAR, $n = 300$, $p = 0.20$ para X_1 e $p = 0.60$ para X_2.

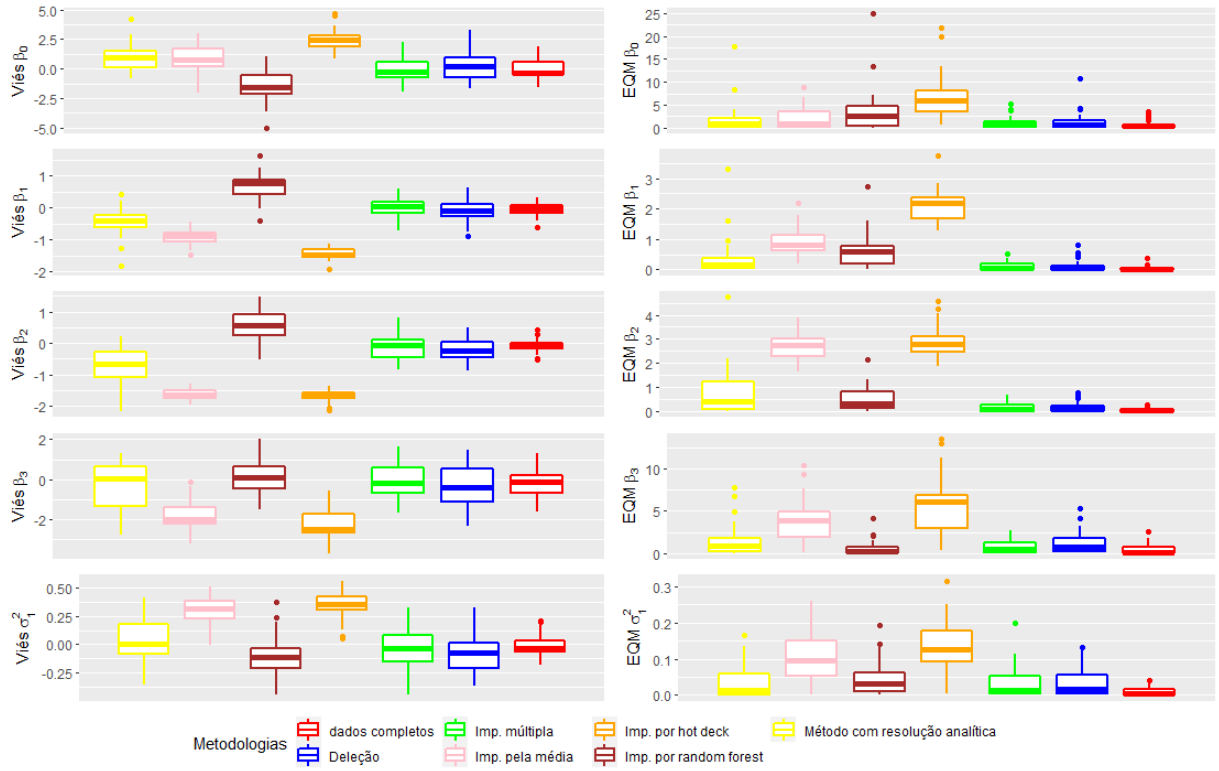

Figura 31 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário com X_1 e X_2 MCAR, $n = 300$, $p = 0.60$ para X_1 e $p = 0.20$ para X_2.

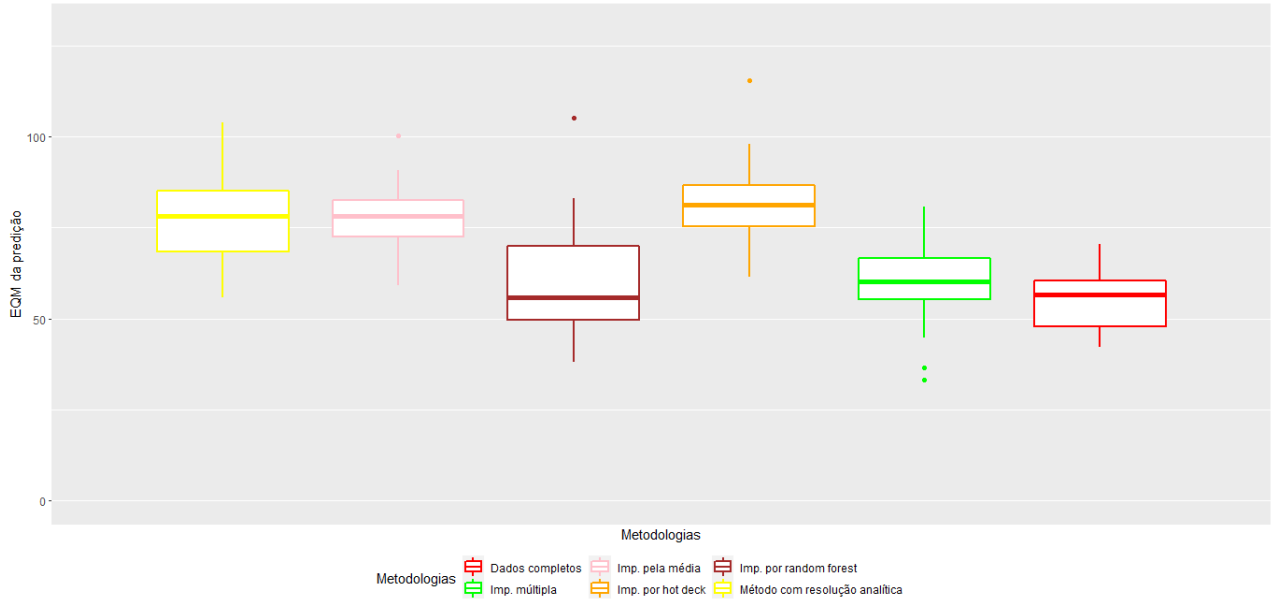

Figura 32 – Erro quadrático médio dos valores preditos para o cenário com X_1 e X_2 MCAR, $n = 300$, $p = 0.60$ para X_1 e $p = 0.20$ para X_2.

Considerando o mecanismo MCAR de geração de dados faltantes para X_1 e X_2, observamos queque o método baseado em modelo aqui proposto, juntamente com o método de Imputação por *Random Forest*, o método de imputação múltipla e o método considerando os dados completos, obtiveram as melhores performances. Para o cenário com duas variáveis faltantes, ressaltamos que o desempenho inferencial e preditivo de todos os métodos comparados são inferiores em relação ao desempenho do modelo estimado com dados completos.

Em relação ao desempenho preditivo nas amostras de teste, o método de imputação por *Random Forest* se sobressai ao método proposto. Já o método de imputação mútipla se apresenta superior ao método baseado em modelo apenas para os cenários com amostras maiores ($n = 300$). Esse bom comportamento preditivo das metodologias de imputação múltipla e imputação por *Random Forest* faz sentido, pois são focadas na predição e o mecanismo de geração dos dados faltantes é o completamente aleatório, quando é esperado que todas as metodologias performem de maneira razoável. Observamos, inclusive, que apesar da imputação por média e *Hot-deck* apresentarem piores desempenhos inferenciais, seus desempenhos preditivos não são muito ruins, especialmente em amostras menores. O método baseado em modelo, em contrapartida, contempla a estimação de 8 parâmetros adicionais (σ_2^2, σ_3^2, σ_4^2, γ_0, γ_1, γ_2, μ_0 e μ_1) que propicia um maior acúmulo de erro na predição desse mecanismo de dados faltantes mais simples.

As Figuras 33, 35, 37 e 39 apresentam os resultados inferenciais dos métodos para o mecanismo MAR de dados faltantes e as Figuras 34, 36, 38 e 40 o desempenho preditivo.

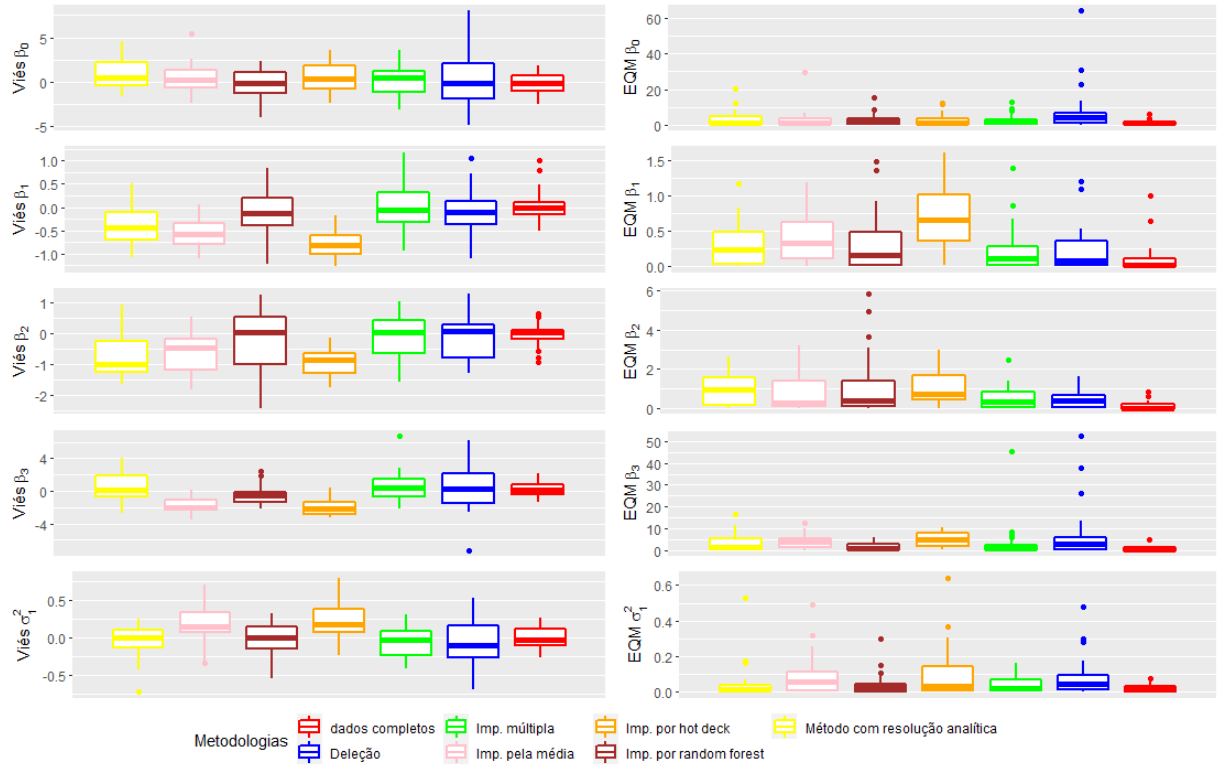

Figura 33 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário com X_1 e X_2 MAR, $n = 100$, $p = 0.20$ para X_1 e $p = 0.60$ para X_2.

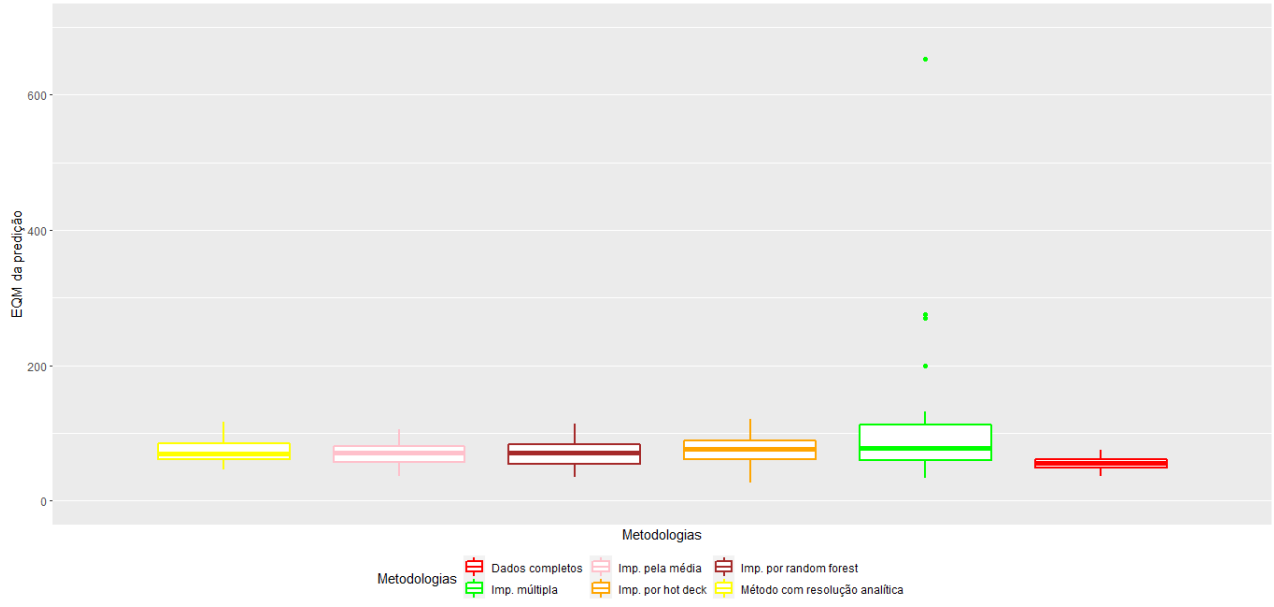

Figura 34 – Erro quadrático médio dos valores preditos para o cenário com X_1 e X_2 MAR, $n = 100$, $p = 0.20$ para X_1 e $p = 0.60$ para X_2.

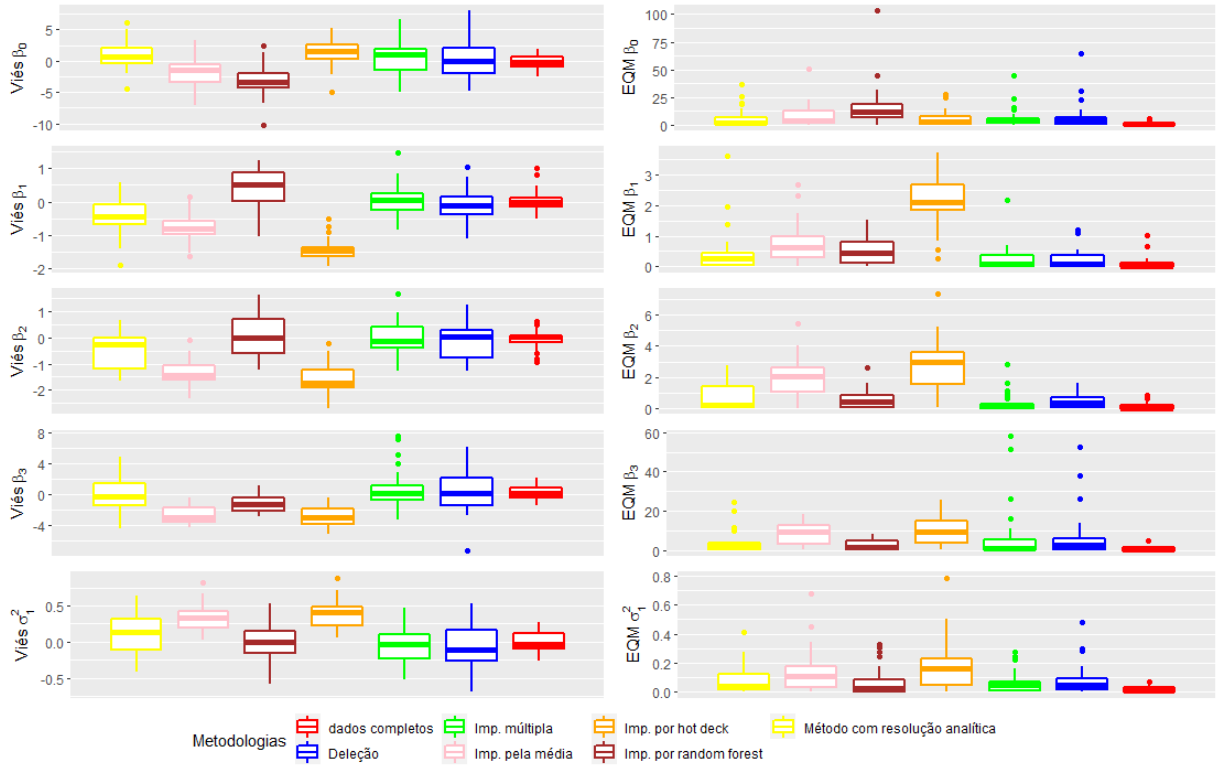

Figura 35 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário com X_1 e X_2 MAR, $n = 100$, $p = 0.60$ para X_1 e $p = 0.20$ para X_2.

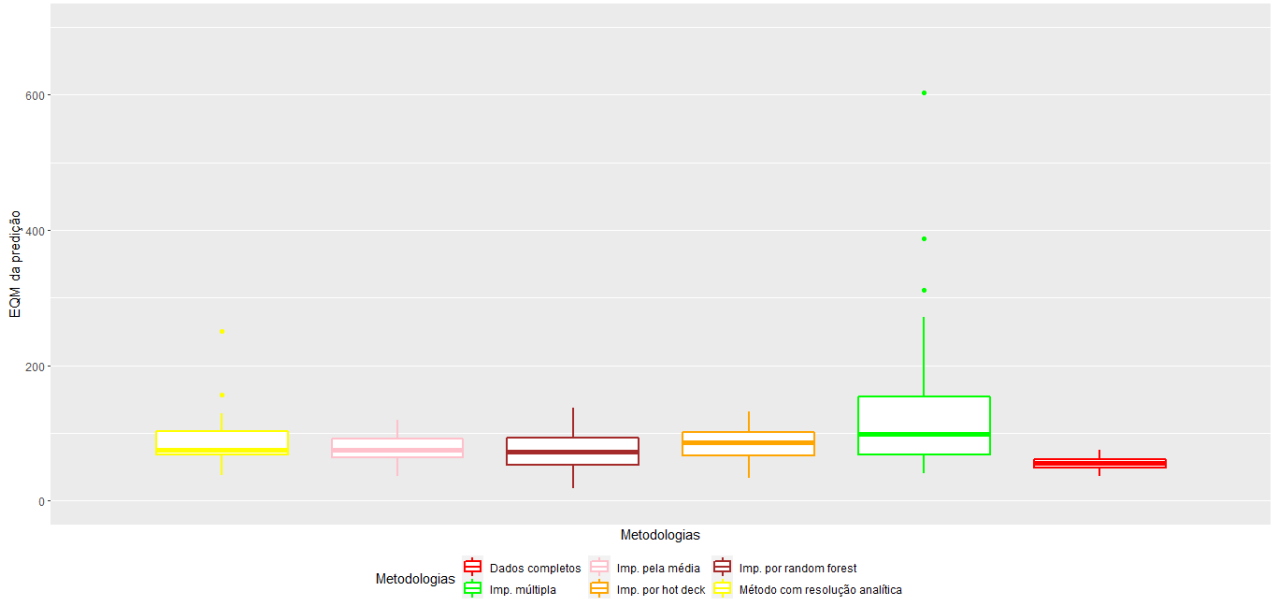

Figura 36 – Erro quadrático médio dos valores preditos para o cenário com X_1 e X_2 MAR, $n = 100$, $p = 0.60$ para X_1 e $p = 0.20$ para X_2.

Para este cenário, na Figura 36, foram detectados mais *outliers*, além dos já exibidos na figura acima, para o método de imputação múltipla, são eles: 1092.81, 2198.80 e 21556.34. Eles não constam no gráfico, pois reduziriam consideralmente a escala dos boxplots e dificultaria, com isso, a análise dos desempenhos dos métodos.

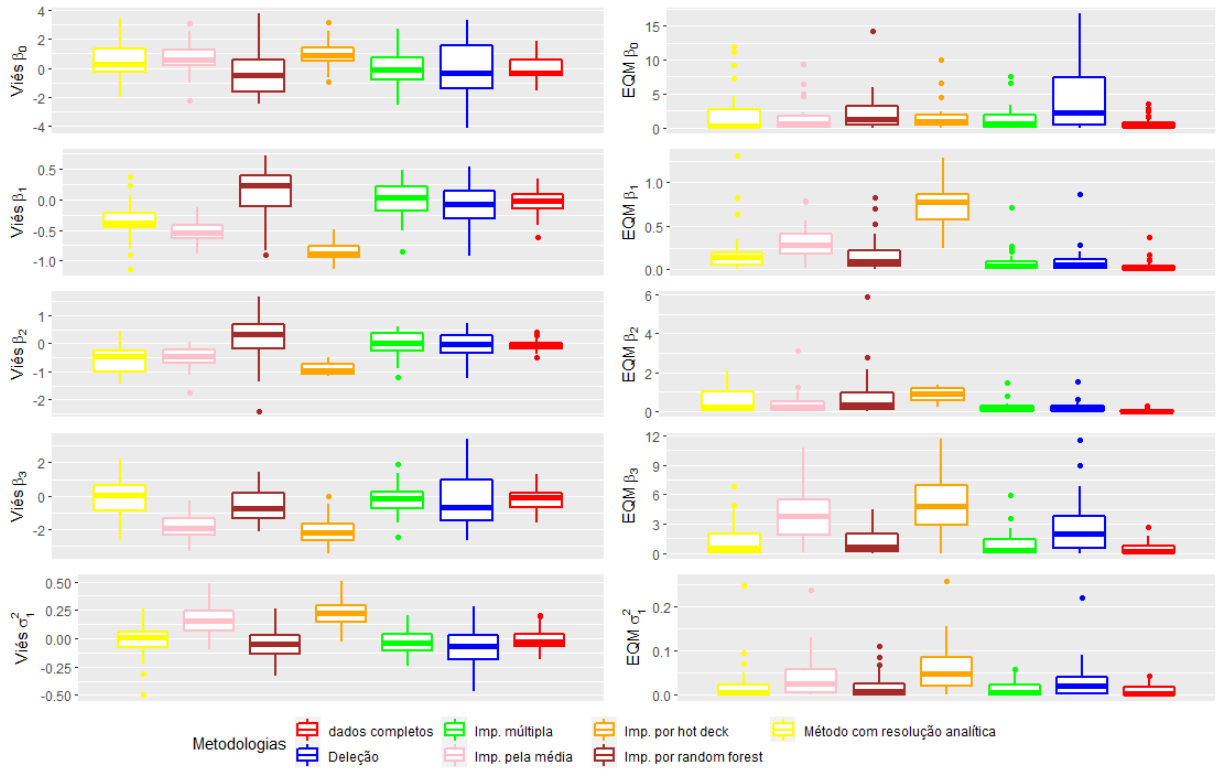

Figura 37 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário com X_1 e X_2 MAR, $n = 300$, $p = 0.20$ para X_1 e $p = 0.60$ para X_2.

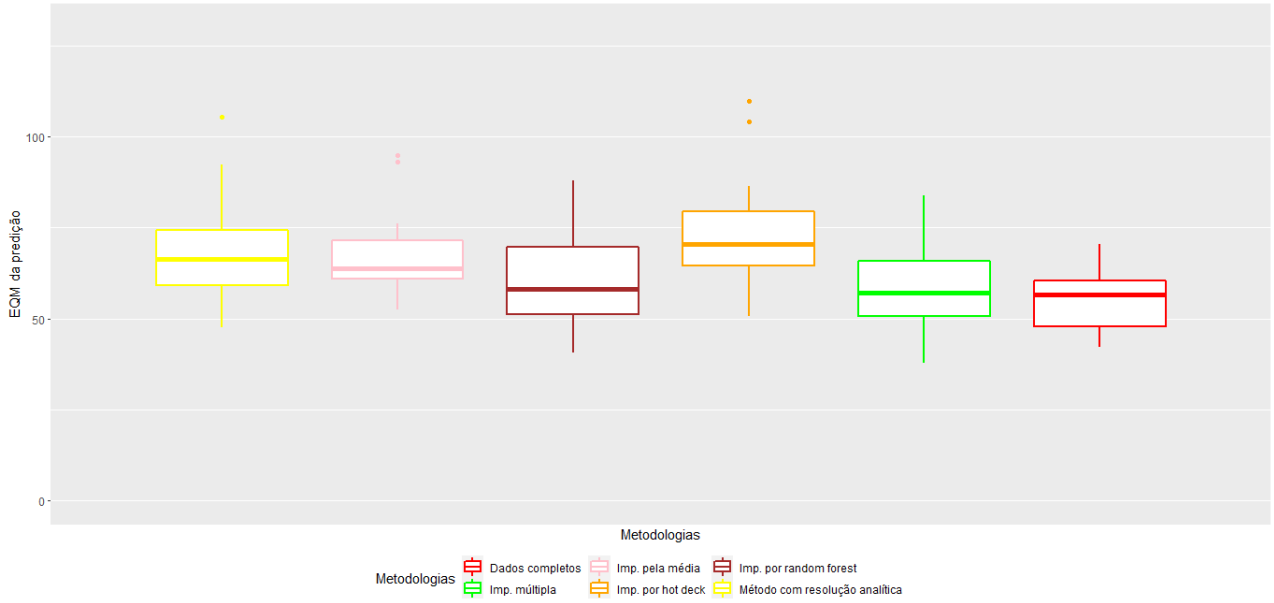

Figura 38 – Erro quadrático médio dos valores preditos para o cenário com X_1 e X_2 MAR, $n = 300$, $p = 0.20$ para X_1 e $p = 0.60$ para X_2.

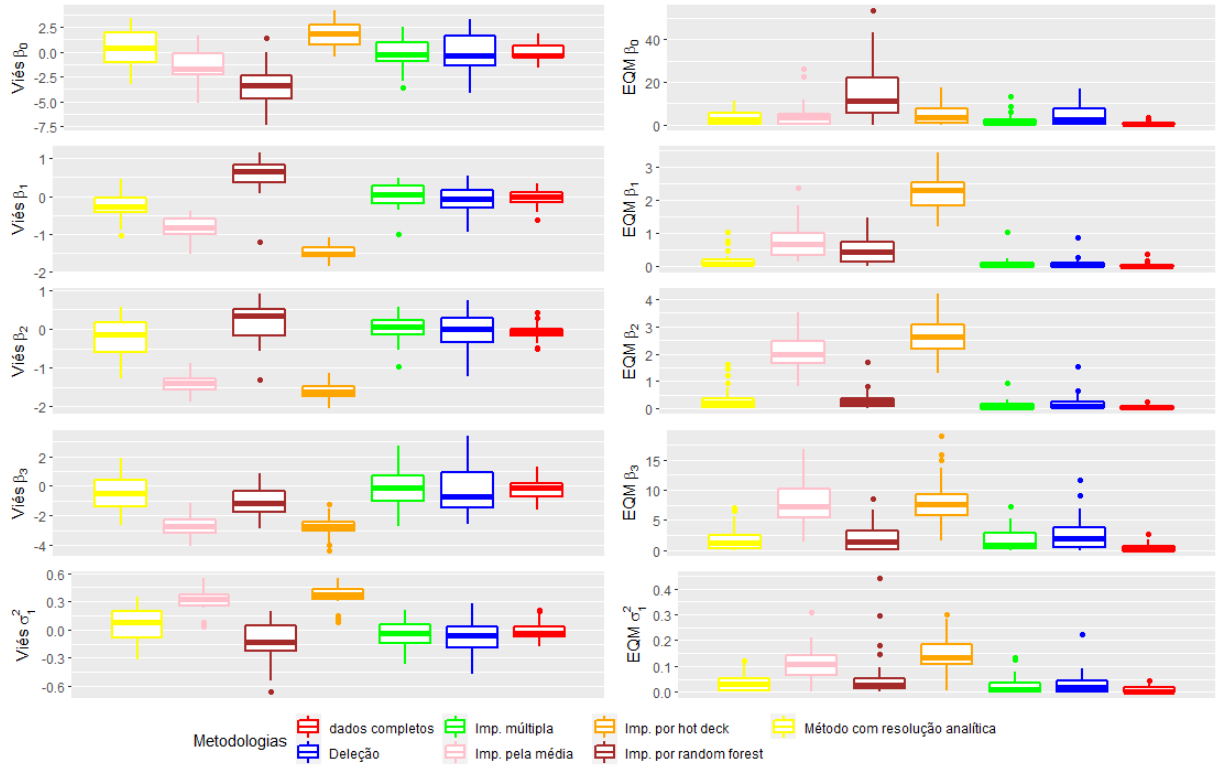

Figura 39 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário com X_1 e X_2 MAR, $n = 300$, $p = 0.60$ para X_1 e $p = 0.20$ para X_2.

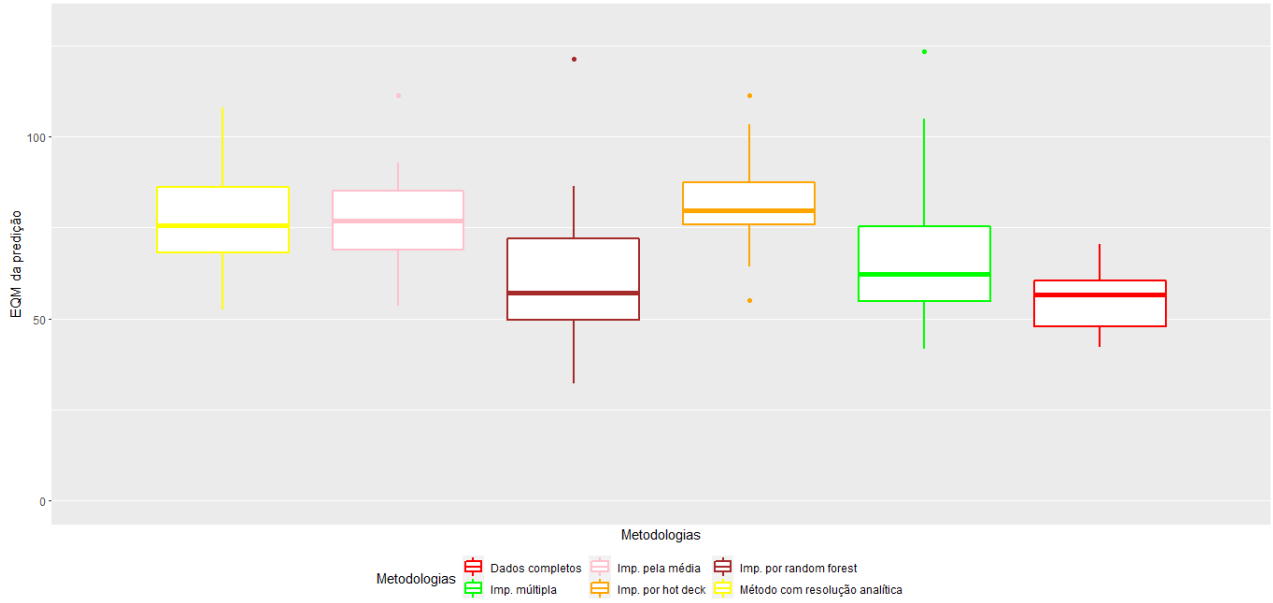

Figura 40 – Erro quadrático médio dos valores preditos para o cenário com X_1 e X_2 MAR, $n = 300$, $p = 0.60$ para X_1 e $p = 0.20$ para X_2.

Para o mecanismo de geração dos dados faltantes MAR, observamos resultados de estimação muito parecidos aos obtidos no método MCAR. O desempenho preditivo da metodologia proposta, para amostras menores ($n = 100$), se aproxima da performance do método de imputação por *Random Forest* e é superior ao método de imputação múltipla, que apresenta também mais valores *outliers*.

As Figuras 41, 43, 45 e 47 apresentam os resultados inferenciais dos métodos para o mecanismo MNAR de dados faltantes e as Figuras 42, 44, 46 e 48 o desempenho preditivo.

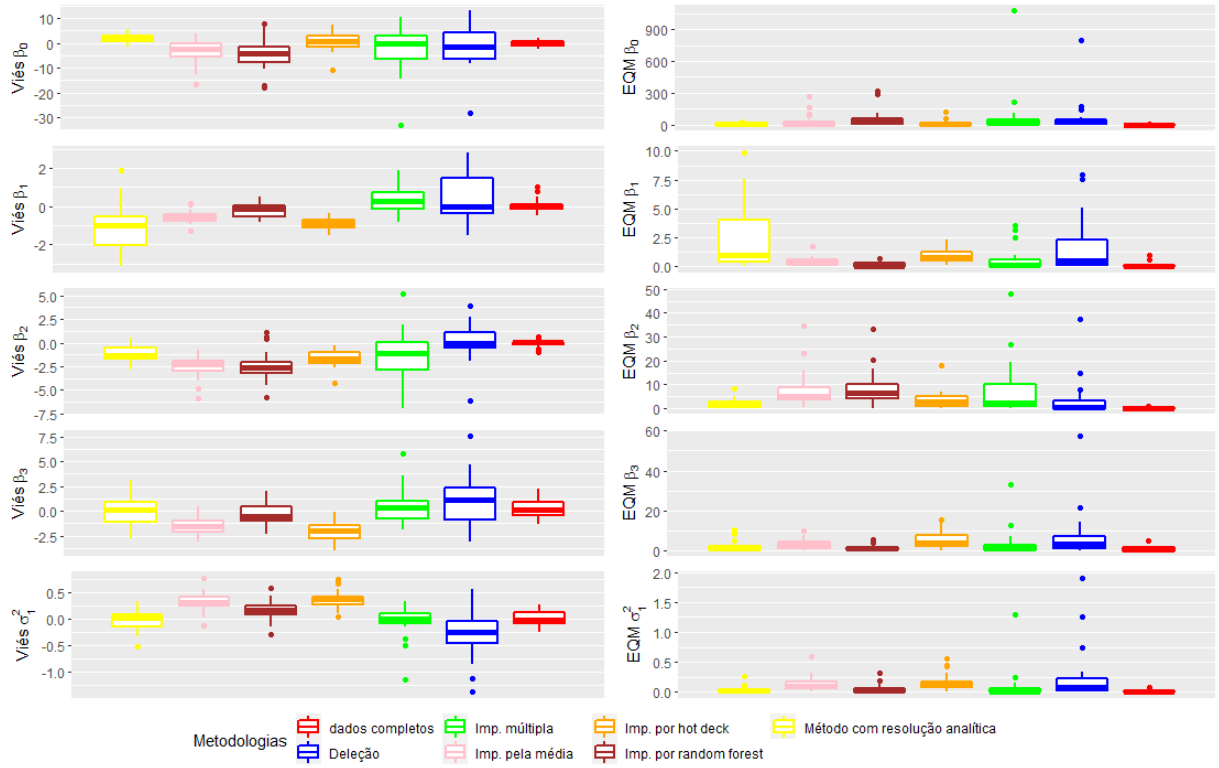

Figura 41 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário com X_1 e X_2 MNAR, $n = 100$, $p = 0.20$ para X_1 e $p = 0.60$ para X_2.

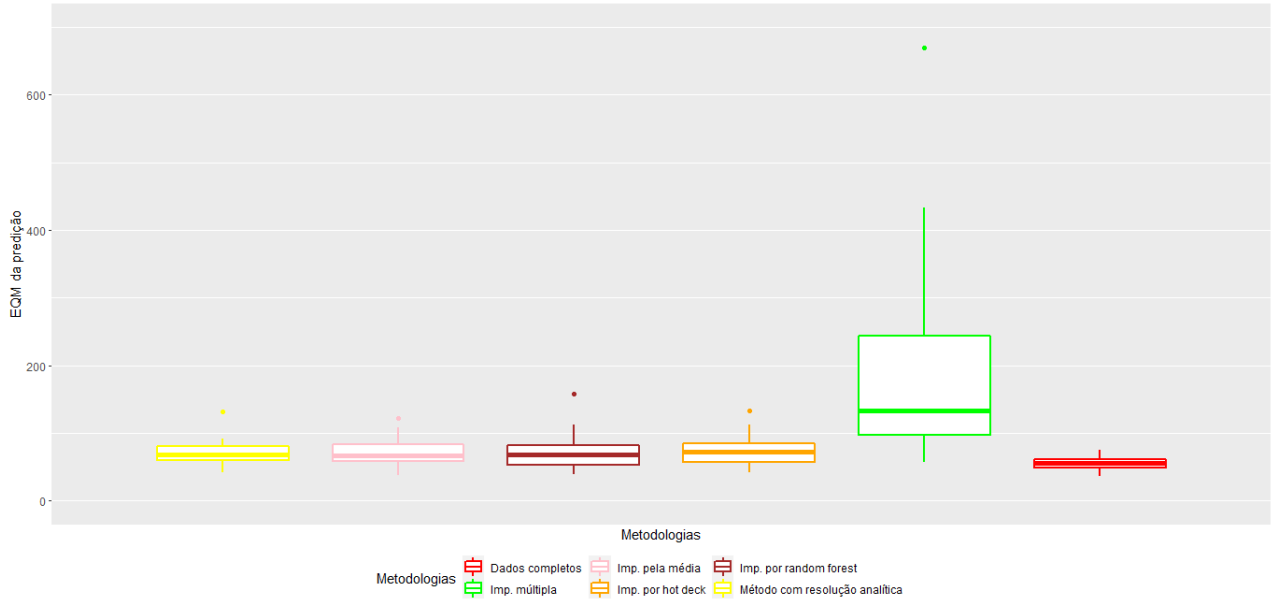

Figura 42 – Erro quadrático médio dos valores preditos para o cenário com X_1 e X_2 MNAR, $n = 100$, $p = 0.20$ para X_1 e $p = 0.60$ para X_2.

Para este cenário, na Figura 42 foram detectados mais *outliers*, além dos já exibidos na figura acima, para o método de imputação múltipla, são eles: 1032.49 e 1320000. Eles não constam no gráfico, pois reduziriam consideralmente a escala dos boxplots e dificultaria, com isso, a análise dos desempenhos dos métodos.

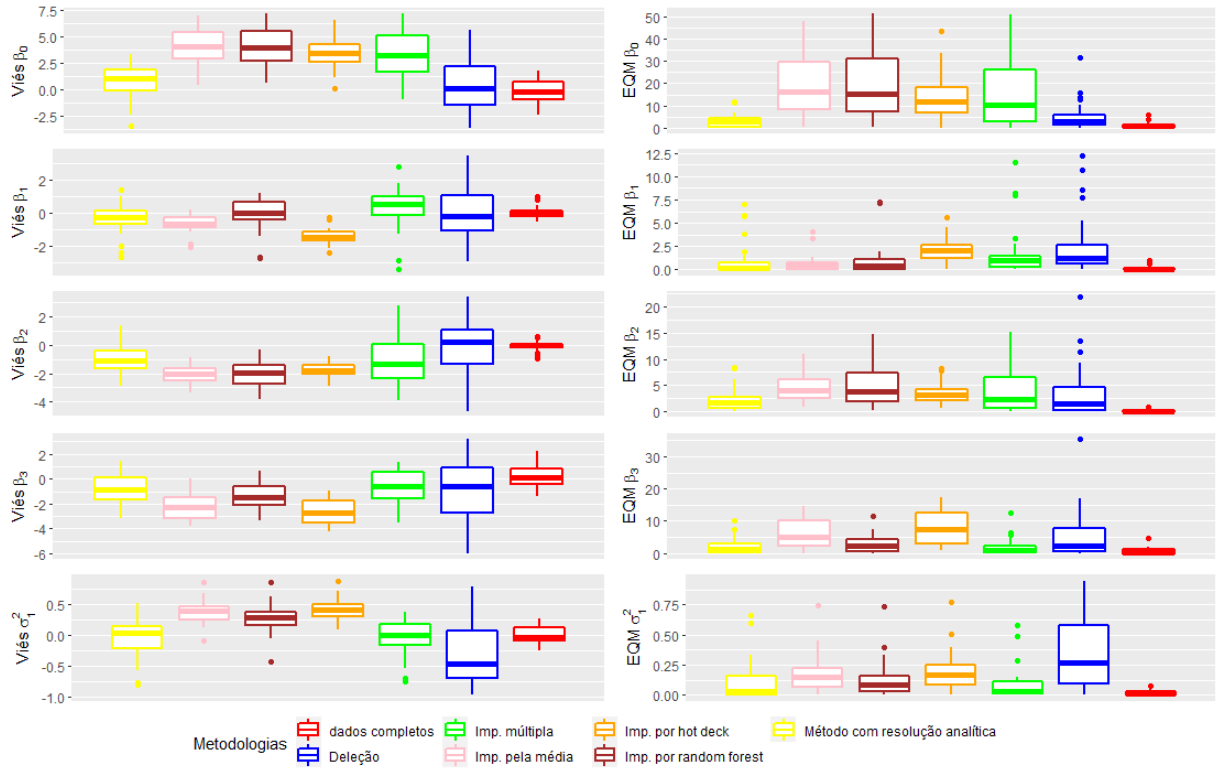

Figura 43 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário com X_1 e X_2 MNAR, $n = 100$, $p = 0.60$ para X_1 e $p = 0.20$ para X_2.

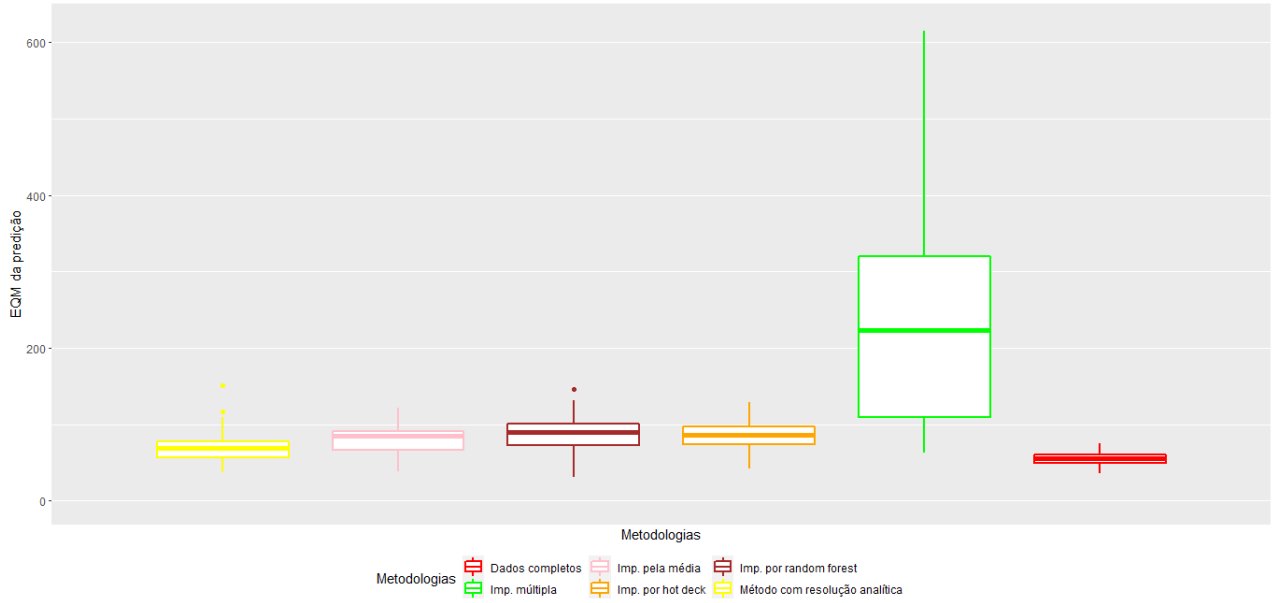

Figura 44 – Erro quadrático médio dos valores preditos para o cenário com X_1 e X_2 MNAR, $n = 100$, $p = 0.60$ para X_1 e $p = 0.20$ para X_2.

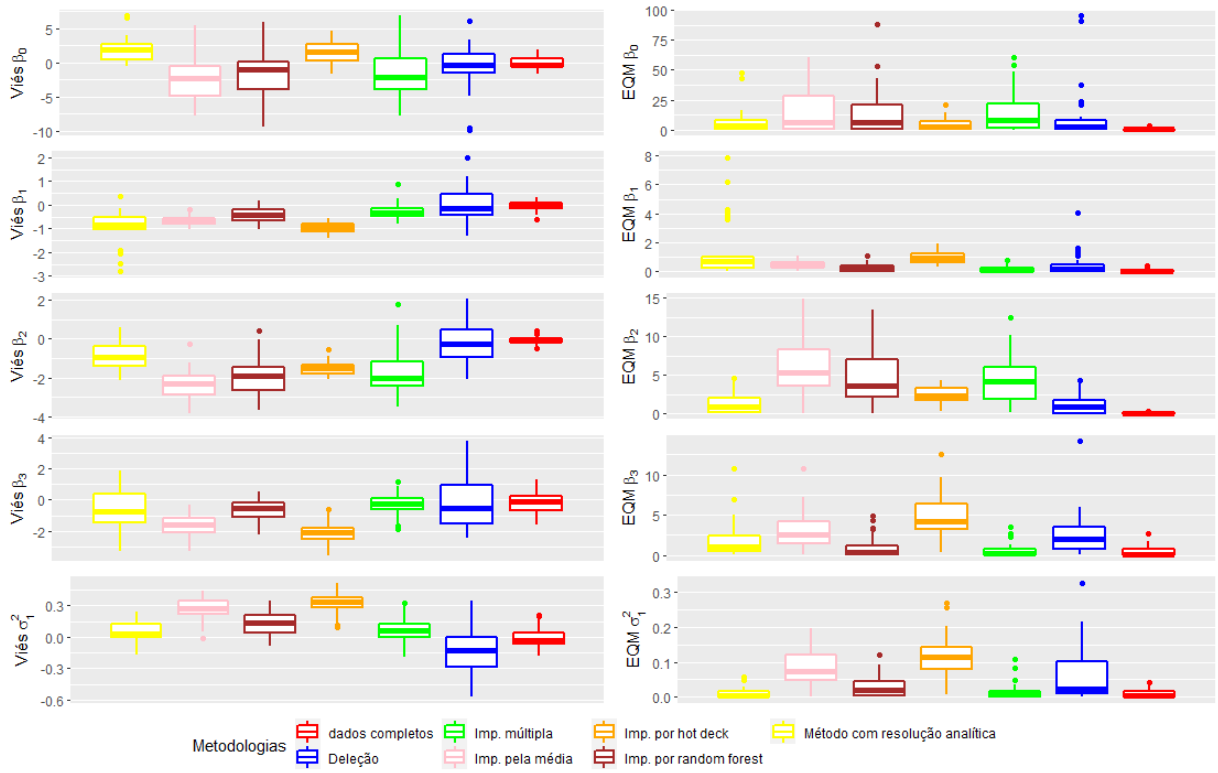

Figura 45 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário com X_1 e X_2 MNAR, $n = 300$, $p = 0.20$ para X_1 e $p = 0.60$ para X_2.

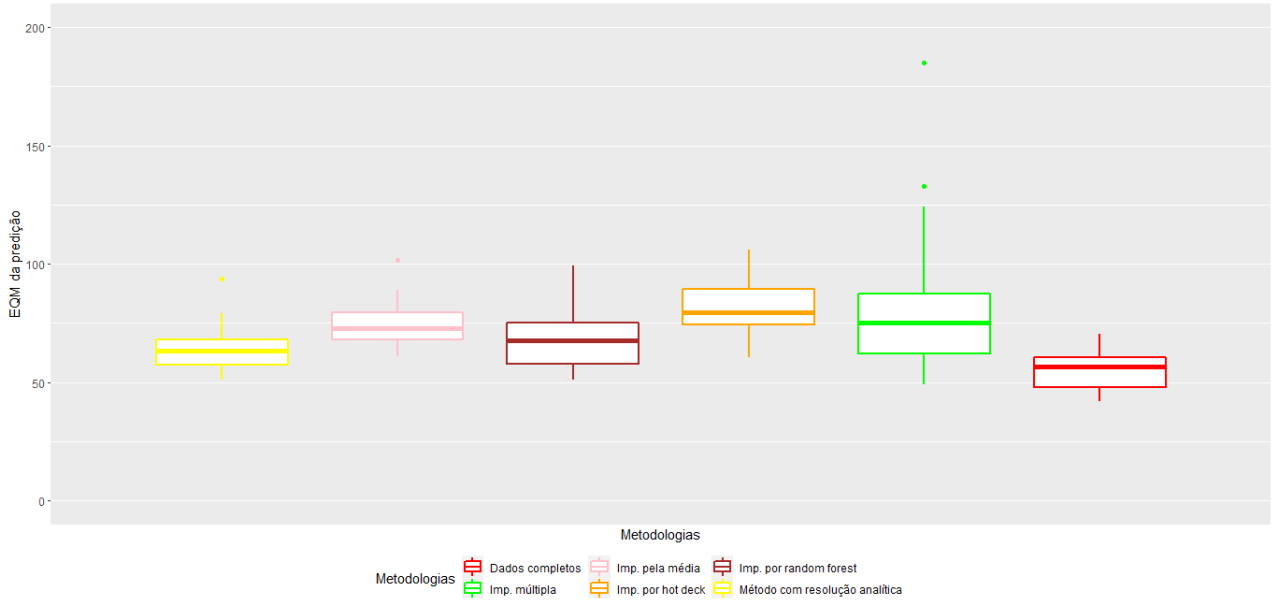

Figura 46 – Erro quadrático médio dos valores preditos para o cenário com X_1 e X_2 MNAR, $n = 300$, $p = 0.20$ para X_1 e $p = 0.60$ para X_2.

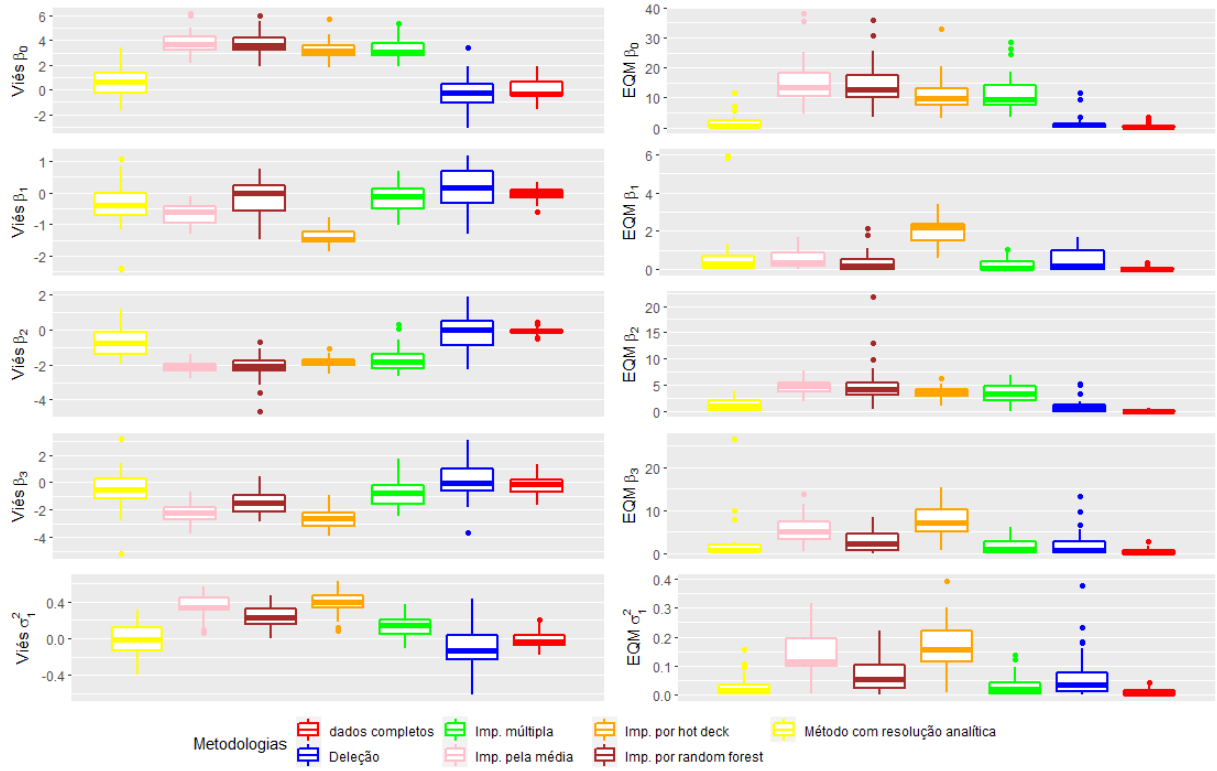

Figura 47 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário com X_1 e X_2 MNAR, $n = 300$, $p = 0.60$ para X_1 e $p = 0.20$ para X_2.

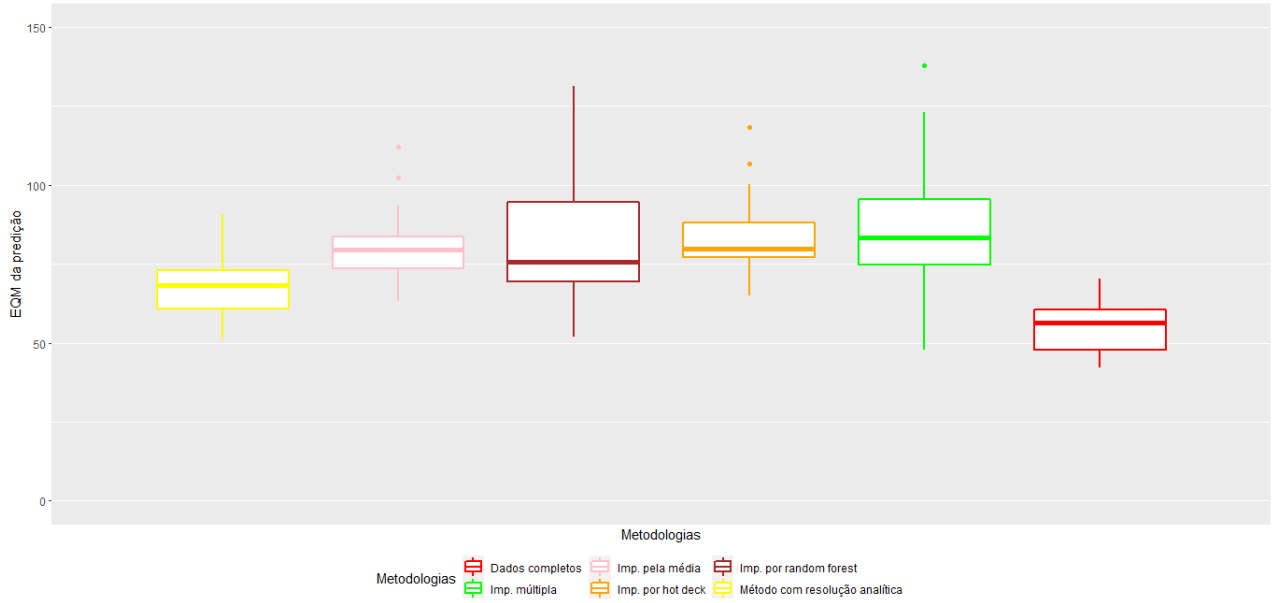

Figura 48 – Erro quadrático médio dos valores preditos para o cenário com X_1 e X_2 MNAR, $n = 300$, $p = 0.60$ para X_1 e $p = 0.20$ para X_2.

Considerando o mecanismo MNAR de geração dos dados faltantes, que provavelmente é a situação de estimação e predição mais desafiadora, o método proposto apresenta, de maneira geral, um desempenho de estimação superior a todos os métodos comparados, exceto à estimação realizada com os dados completos, mesmo contemplando a estimação de 8 parâmetros adicionais. O método de estimação baseado em modelo aqui proposto também apresenta resultados preditivos superiores a todos os métodos, exceto ao modelo estimado com dados completos.

CAPÍTULO

5

MÉTODOS BASEADOS EM MODELO SEM RESOLUÇÃO ANALÍTICA

Neste capítulo, desenvolvemos a segunda proposta desta tese, que são métodos para estimação de modelos na presença de dados faltantes quando a solução analítica não existe. Aplicamos os métodos em modelo de regressão linear múltipla com k variáveis, em que consideramos, primeiramente, apenas X_1 com valores faltantes e, depois, X_1 e X_2 apresentando informações incompletas. Esse caso se aplica quando assumimos distribuições não normais para a variável resposta e/ou faltantes, ou outras distribuições para as quais a resolução analítica não se aplica.

5.1 Modelo com uma variável faltante

Sejam Y, X_1, X_2 variáveis aleatórias, ou seja, $k = 2$. Então:

$$f(y|x_1, x_2, \theta) = \frac{f(y, x_1, x_2, \theta)}{f(x_1, x_2, \theta)}. \tag{5.1}$$

Da Equação (5.1) temos:

$$f(y, x_1, x_2, \theta) = f(y|x_1, x_2, \theta) f(x_1, x_2, \theta). \tag{5.2}$$

Sendo assim, como vimos nos cálculos da Seção 4.1:

$$f(y, x_1|x_2, \theta) = f(y|x_1, x_2, \theta) f(x_1|x_2, \theta). \tag{5.3}$$

Veja que estamos admitindo, neste caso, que X_1 possui valores faltantes, logo consideramos a distribuição de $f(y|x_2, \theta)$ para quando isso acontece, que é definida como:

$$f(y|x_2,\theta) = \int f(y|x_1,x_2,\theta) f(x_1|x_2,\theta) dx_1. \tag{5.4}$$

Dadas as distribuições de $Y|X_1,X_2$ e $X_1|X_2$, nesta seção assumimos que a integral (5.4) não possui resolução analítica. Neste capítulo, novamente, diferente de um modelo de regressão linear tradicional, cujas variáveis explicativas são vistas como características fixas, as variáveis explicativas com valores faltantes são analisadas pelos métodos aqui desenvolvidos como variáveis aleatórias para as quais definimos uma distribuição de probabilidade de acordo com suas naturezas. A esperança das variáveis explicativas com valores faltantes são então definidas como função das outras variáveis explicativas disponíveis. Apresentamos, a seguir, três soluções basedas em modelo para este cenário: método por integração numérica, método utilizando média de log-verossimilhanças e método utilizando o algoritmo EM.

5.1.1 Método por integração numérica

Neste método, utilizamos uma aproximação Monte Carlo para encontrarmos a densidade $f(y|x_2,\theta)$, que é expressa pela integral (5.4). Sabemos que:

$$\begin{aligned} f(y|x_2,\theta) &= \int f(y|x_1,x_2,\theta) f(x_1|x_2,\theta) dx_1 \\ &= E_{X_1|X_2}[f(y|X_1,x_2,\theta)]. \end{aligned} \tag{5.5}$$

Seja i uma observação que possui valor faltante em X_1. Então, geramos M valores aleatórios para esta observação, $x_{1i1},...,x_{1iM}$, da distribuição de $X_1|X_2$ e a aproximação Monte Carlo de (5.5), para esta observação, se torna:

$$\begin{aligned} f(y_i|x_{2i},\theta) &= E_{X_1|X_2}[f(y_i|X_{1i},x_{2i},\theta)] \\ &= \frac{1}{M}\sum_{m=1}^{M} f(y_i|x_{1im},x_{2i},\theta). \end{aligned} \tag{5.6}$$

Consideremos, agora, $\delta_1(i)$ como a função indicadora de quando uma observação não é faltante em relação a X_1, conforme descrita na Seção 4.1. A função de verossimilhança completa, considerando estes dois cenários possíveis (i ser observada ou i ser faltante), pode ser escrita como:

$$\begin{aligned} L(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2) = \prod_{i=1}^{n} & \Big[f(y_i|x_{2i},\theta)^{1-\delta_1(i)} \\ & \times \quad (f(y_i|x_{1i},x_{2i},\theta) f(x_{1i}|x_{2i},\theta))^{\delta_1(i)} \Big] \end{aligned} \tag{5.7}$$

em que θ é o vetor de parâmetros a ser estimado.

Observe que, na Equação (5.7), utilizamos $f(y_i, x_{1i}|x_{2i}, \theta)$ e não $f(y_i|x_{1i}, x_{2i}, \theta)$ para as observações completas. Realizamos este procedimento para que o algoritmo estime os parâmetros da distribuição de $X_1|X_2$ para os casos completos (sem informações faltantes) e, com isso, possa melhorar a estimativa dos parâmetros para os casos com valores faltantes que dependem da geração de amostras desta distribuição.

Nosso objetivo é encontrar as estimativas que tornam máximo o valor da função de verossimilhança completa em (5.7). Dessa forma, o próximo passo para encontrarmos as estimativas de máxima-verossimilhança de θ é calcularmos a função de log-verossimilhança $l(\theta|\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2)$ completa correspondente a (5.7):

$$\begin{aligned} l(\theta|\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2) &= \log L(\theta|\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2) \\ &= \sum_{i=1}^{n} (1 - \delta_1(i)) (\log f(y_i|x_{2i}, \theta)) \\ &+ \delta_1(i) (\log[f(y_i|x_{1i}, x_{2i}, \theta) f(x_{1i}|x_{2i}, \theta)]) . \end{aligned} \tag{5.8}$$

De acordo com a Equação (5.6), vimos que a densidade $f(y_i|x_{2i}, \theta)$ pode ser aproximada por Monte Carlo. Logo, a função de log-verossimilhança completa dada em (5.8), pode ser reescrita como:

$$\begin{aligned} l(\theta|\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2) &= \sum_{i=1}^{n} (1 - \delta_1(i)) (\log f(y_i|x_{2i}, \theta)) \\ &+ \delta_1(i) (\log[f(y_i|x_{1i}, x_{2i}, \theta) f(x_{1i}|x_{2i}, \theta)]) \\ &\approx \sum_{i=1}^{n} (1 - \delta_1(i)) \left(\log \left(\frac{1}{M} \sum_{m=1}^{M} f(y_i|x_{1im}, x_{2i}, \theta) \right) \right) \\ &+ \delta_1(i) (\log[f(y_i|x_{1i}, x_{2i}, \theta) f(x_{1i}|x_{2i}, \theta)]) \\ &= \sum_{i=1}^{n} (1 - \delta_1(i)) \left(\log \left(\frac{1}{M} \sum_{m=1}^{M} f(y_i|x_{1im}, x_{2i}, \theta) \right) \right) \\ &+ \delta_1(i) ([\log(f(y_i|x_{1i}, x_{2i}, \theta)) + \log(f(x_{1i}|x_{2i}, \theta))]), \end{aligned} \tag{5.9}$$

sendo $x_{1i1}, \ldots, x_{1iM}$ os M valores aleatórios gerados da distribuição de $X_1|X_2$, para uma observação i com valor faltante em X_1.

5.1.2 Método utilizando média de log-verossimilhanças

Neste método, maximizamos uma média de log-verossimilhanças. Para isto, consideramos $L(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2)$ a função de verossimilhança completa, que é dada por:

$$L(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2) = \prod_{i=1}^{n} f(y_i,x_{1i}|x_{2i},\theta), \tag{5.10}$$

ou em função das distribuições condicionais,

$$L(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2) = \prod_{i=1}^{n} f(y_i|x_{1i},x_{2i},\theta) f(x_{1i}|x_{2i},\theta). \tag{5.11}$$

O logaritmo da função de verossimilhança completa $l(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2) = \log L(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2)$ é dado por:

$$l(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2) = \sum_{i=1}^{n} \log[f(y_i|x_{1i},x_{2i},\theta) f(x_{1i}|x_{2i},\theta)]. \tag{5.12}$$

Seja n_2 o número de observações com valores faltantes em X_1 $(n_2 < n)$. Para estes casos, geramos M valores aleatórios $x_{1i1},...,x_{1iM}$ com $i = 1,2,...,n_2$, da distribuição de $X_1|X_2$. Então, uma aproximação de (5.12), $\bar{l}(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2)$, para as observações que possuem valores faltantes em X_1, é:

$$\begin{aligned} \bar{l}(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2) &\approx \frac{1}{M}\sum_{m=1}^{M}\left[\sum_{i=1}^{n_2}\log[f(y_i|x_{1im},x_{2i},\theta) f(x_{1im}|x_{2i},\theta)]\right] \\ &= \frac{1}{M}\sum_{m=1}^{M}\left[\sum_{i=1}^{n_2}[\log(f(y_i|x_{1im},x_{2i},\theta)) + \log(f(x_{1im}|x_{2i},\theta))]\right]. \end{aligned} \tag{5.13}$$

Considerando agora a função indicadora de quando uma observação i não é faltante em relação a X_1, $\delta_1(i)$, podemos estender o cálculo de $\bar{l}(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2)$ para todas as observações pertencentes aos dois cenários possíveis (i com valor faltante em X_1 e i com valor observado em X_1) da seguinte forma:

$$\begin{aligned} \bar{l}(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2) \approx \frac{1}{M}\sum_{m=1}^{M} &\left[\sum_{i=1}^{n}(1-\delta_1(i))\left([\log(f(y_i|x_{1im},x_{2i},\theta)) + \log(f(x_{1im}|x_{2i},\theta))]\right)\right. \\ &+ \left.\delta_1(i)\left([\log(f(y_i|x_{1i},x_{2i},\theta)) + \log(f(x_{1i}|x_{2i},\theta))]\right)\right]. \end{aligned} \tag{5.14}$$

5.1.3 Método utilizando o algoritmo EM

Neste método, resolvemos o problema da integral (5.4) não possuir resolução analítica, utilizando o algoritmo EM-Monte Carlo para a maximização da função de verossimilhança. Para

o algoritmo EM, precisamos do logaritmo da função de verossimilhança completa $l(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2) = \log L(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2)$ que, conforme construído na Seção 5.1.2, é dado por:

$$l(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2) = \sum_{i=1}^{n} \log[f(y_i|x_{1i},x_{2i},\theta)f(x_{1i}|x_{2i},\theta)]. \tag{5.15}$$

A cada iteração do algoritmo EM alternam-se um passo E e um passo M. Seja θ_r o valor de θ da r-ésima iteração. O $(r+1)$-ésimo passo E, considerando X_1 faltante, consiste em calcular:

$$\begin{aligned} Q(\theta|\theta_r) &= E_{\mathbf{X}_1|\mathbf{Y},\mathbf{X}_2,\theta_r}[l(\theta|\mathbf{Y},\mathbf{X}_1,\mathbf{X}_2)|\mathbf{Y}=\mathbf{y},\mathbf{X}_2=\mathbf{x}_2,\theta_r] \\ &= E_{\mathbf{X}_1|\mathbf{Y},\mathbf{X}_2,\theta_r}\left(\sum_{i=1}^{n}\log[f(y_i|x_{1i},x_{2i},\theta)f(x_{1i}|x_{2i},\theta)]\right) \\ &= \sum_{i=1}^{n} E_{\mathbf{X}_1|\mathbf{Y},\mathbf{X}_2,\theta_r}[\log(f(y_i|x_{1i},x_{2i},\theta)f(x_{1i}|x_{2i},\theta))] \\ &= \sum_{i=1}^{n} q_i(\theta|\theta_r). \end{aligned} \tag{5.16}$$

O passo M realiza a maximização de (5.16) com respeito a θ, resultando em uma nova estimativa θ_{r+1}. Dado um ponto θ_0, a iteração entre o passo E e o passo M é repetida até a convergência.

Para os casos em que a esperança em (5.16) não possui forma analítica, ela pode ser estimada por aproximações de Monte Carlo. Logo, iremos gerar M valores aleatórios para X_1, $x_{1i1}^{(r)},...,x_{1iM}^{(r)}$ para $i = 1,2,...,n_2$, sendo n_2 o número de observações com valores faltantes em X_1 $(n_2 < n)$, da distribuição de $X_1|Y,X_2$. Então, a aproximação de Monte Carlo para q_i, quando uma observação i possuir valor faltante em X_1, é:

$$\begin{aligned} q_i(\theta|\theta_r) &= E_{\mathbf{X}_1|\mathbf{Y},\mathbf{X}_2,\theta_r}[l(\theta|Y_i,X_{1i},X_{2i})|y_i,x_{2i},\theta_r] \\ &\approx \frac{1}{M}\sum_{m=1}^{M} l(\theta|y_i,x_{1im}^{(r)},x_{2i}) \\ &= \frac{1}{M}\sum_{m=1}^{M}\left[\log[f(y_i|x_{1im}^{(r)},x_{2i},\theta_r)f(x_{1im}^{(r)}|x_{2i},\theta_r)]\right] \\ &= \frac{1}{M}\sum_{m=1}^{M}\left[\log(f(y_i|x_{1im}^{(r)},x_{2i},\theta_r)) + \log(f(x_{1im}^{(r)}|x_{2i},\theta_r))\right]. \end{aligned} \tag{5.17}$$

Como precisamos gerar M valores aleatórios para X_1 da distribuição de $X_1|Y,X_2$ para o cálculo de $q_i(\theta|\theta_r)$ e esta distribuição não é tradicional, precisamos recorrer à utilização de

algum método de simulação. Nesse trabalho, vamos utilizar o método de amostragem por importância (*importance sampling*), utilizando a densidade $f(x_1|x_2,\theta_r)$. A versão por amostragem de importância de (5.17) é:

$$\begin{aligned}
q_i(\theta|\theta_r) &= E_{\mathbf{X}_1|\mathbf{Y},\mathbf{X}_2,\theta_r}[l(\theta|Y_i,X_{1i},X_{2i})|y_i,x_{2i},\theta_r] \\
&= \int l(\theta|y_i,x_{1i},x_{2i}) f(x_{1i}|y_i,x_{2i},\theta_r) dx_{1i} \\
&= \int l(\theta|y_i,x_{1i},x_{2i}) \frac{f(x_{1i}|y_i,x_{2i},\theta_r)}{f(x_{1i}|x_{2i},\theta_r)} f(x_{1i}|x_{2i},\theta_r) dx_{1i} \\
&= E_{X_1|X_2}\left[l(\theta|y_i,X_{1i},x_{2i}) \frac{f(X_{1i}|y_i,x_{2i},\theta_r)}{f(X_{1i}|x_{2i},\theta_r)} \right] \\
&\approx \frac{1}{M}\sum_{m=1}^{M}\left[l(\theta|y_i,x_{1im}^{(r)},x_{2i}) \frac{f(x_{1im}^{(r)}|y_i,x_{2i},\theta_r)}{f(x_{1im}^{(r)}|x_{2i},\theta_r)} \right] \\
&= \frac{1}{M}\sum_{m=1}^{M}\left[k_{rim}\, l(\theta|y_i,x_{1im},x_{2i}) \right] \\
&= \frac{1}{M}\sum_{m=1}^{M}\left[k_{rim}\left[\log(f(y_i|x_{1im}^{(r)},x_{2i},\theta_r)) + \log(f(x_{1im}^{(r)}|x_{2i},\theta_r)) \right] \right] \qquad (5.18)
\end{aligned}$$

em que $k_{rim} = \frac{f(x_{1im}^{(r)}|y_i,x_{2i},\theta_r)}{f(x_{1im}^{(r)}|x_{2i},\theta_r)}$ são os pesos da amostragem por importância e $x_{1i1}^{(r)},...,x_{1iM}^{(r)}$, para $i = 1,2,...,n_2$, são agora gerados da distribuição conhecida de $X_1|X_2$.

A expressão k_{rim} pode ainda ser simplificada como:

$$\begin{aligned}
k_{rim} &= \frac{f(x_{1im}^{(r)}|y_i,x_{2i},\theta_r)}{f(x_{1im}^{(r)}|x_{2i},\theta_r)} = \frac{f(x_{1im}^{(r)},y_i,x_{2i}|\theta_r)}{f(y_i,x_{2i}|\theta_r) f(x_{1im}^{(r)}|x_{2i},\theta_r)} \\
&= \frac{f(y_i|x_{1im}^{(r)},x_{2i},\theta_r) f(x_{1im}^{(r)},x_{2i}|\theta_r)}{f(y_i|x_{2i},\theta_r) f(x_{2i}|\theta_r) f(x_{1im}^{(r)}|x_{2i},\theta_r)} \\
&= \frac{f(y_i|x_{1im}^{(r)},x_{2i},\theta_r) f(x_{1im}^{(r)}|x_{2i},\theta_r) f(x_{2i}|\theta_r)}{f(y_i|x_{2i},\theta_r) f(x_{2i}|\theta_r) f(x_{1im}^{(r)}|x_{2i},\theta_r)} \\
&= \frac{f(y_i|x_{1im}^{(r)},x_{2i},\theta_r)}{f(y_i|x_{2i},\theta_r)} \\
&= \frac{f(y_i|x_{1im}^{(r)},x_{2i},\theta_r)}{\int f(y_i,x_{1i}|x_{2i},\theta_r) dx_{1i}} \\
&= \frac{f(y_i|x_{1im}^{(r)},x_{2i},\theta_r)}{\int f(y_i|x_{1i},x_{2i},\theta_r) f(x_{1i}|x_{2i},\theta_r) dx_{1i}}. \qquad (5.19)
\end{aligned}$$

O peso k_{rim} pode ser aproximado por:

$$k_{rim} \approx \frac{f(y_i|x_{1im}^{(r)}, x_{2i}, \theta_r)}{\frac{1}{M}\sum_{m=1}^{M} f(y_i|x_{1im}^{(r)}, x_{2i}, \theta_r)}. \tag{5.20}$$

Considerando agora $\delta_1(i)$ a função indicadora de quando uma observação i não é faltante em relação a X_1, o cálculo de $Q(\theta|\theta_r)$ para todas as observações pertencentes aos dois cenários possíveis (i com valor faltante em X_1 e i com valor observado em X_1), se torna:

$$\begin{aligned} Q(\theta|\theta_r) \approx \frac{1}{M}\sum_{m=1}^{M} & \left[\sum_{i=1}^{n}(1-\delta_1(i))\left(k_{rim}\left[\log(f(y_i|x_{1im}^{(r)}, x_{2i}, \theta_r)) + \log(f(x_{1im}^{(r)}|x_{2i}, \theta_r))\right]\right)\right. \\ + & \left.\delta_1(i)\left([\log(f(y_i|x_{1i}, x_{2i}, \theta_r)) + \log(f(x_{1i}|x_{2i}, \theta_r))]\right)\right]. \end{aligned} \tag{5.21}$$

5.1.4 Análise preditiva dos métodos

Para analisarmos o desempenho preditivo dos métodos propostos, dividimos o banco de dados em 70% para treino (subconjunto através do qual estimamos os parâmetros) e 30% para teste (subconjunto para o qual calculamos os valores preditos $\widehat{y}$ e comparamos com os observados y). Para o cálculo do $\widehat{y}$ para as observações na base de teste, consideramos o valor esperado estimado da distribuição de $Y|X_1, X_2$. Logo,

i) Se x_{1i} é observado, temos:

$$\widehat{y_i} = E[Y_i|x_{1i}, x_{2i}, \widehat{\theta}] \tag{5.22}$$

em que $\widehat{\theta}$ é o vetor das estimativas dos parâmetros;

ii) Se x_{1i} é faltante, temos:

$$\widehat{y_i} = \frac{1}{M}\sum_{m=1}^{M} E[Y_i|x_{1im}, x_{2i}, \widehat{\theta}] \tag{5.23}$$

em que $\widehat{\theta}$ é o vetor das estimativas dos parâmetros e, neste cenário, são gerados M valores aleatórios para X_{1i} da distribuição de $X_1|X_2$.

Efetuado o cálculo do $\widehat{y}$ para a base de teste, calculamos a média das diferenças ao quadrado entre $\widehat{y}$ e y.

5.1.5 Estudo de simulação

Nesta seção, será discutido como foi feito o estudo de simulação e comparação do desempenho dos métodos propostos para a abordagem sem resolução analítica com o de outros métodos de imputação e deleção de dados. Nesse cenário com apenas uma variável com valores

faltantes, comparamos o desempenho dos diferentes métodos em relação ao viés e ao erro quadrático médio (EQM) das estimativas dos parâmetros e em relação ao poder preditivo definido pela média das diferenças quadráticas entre $\widehat{y}$ e y em amostras teste, separadas especificamente para esse fim.

Vários cenários de simulação foram testados, entre os quais variamos: o tamanho da amostra ($n = 100$ e $n = 300$), a proporção de valores faltantes ($p = 0.20$ e $p = 0.60$) e o mecanismo que gera os dados faltantes (MCAR, MAR e MNAR). Para cada cenário analisado, simulamos 30 réplicas (amostras diferentes) sob as mesmas condições. Com elas, temos amostras de tamanho 30 para conduzir análises de desempenho através do viés e EQM das estimativas dos parâmetros, assim como do erro de predição. Quanto mais próximas de zero estiverem os valores dessas métricas, mais precisa é a estimação e a predição. Aqui, adotamos o modelo Weibull como a distribuição de probabilidades das variáveis, mas qualquer outro modelo pode ser considerado e a metodologia adaptada.

Para cada conjunto de dados, separamos as 70% primeiras observações para treino e estimação, através da qual fazemos a análise inferencial dos parâmetros e os outros 30% para teste, em que analisamos os métodos propostos em relação ao poder preditivo, comparando-o com os métodos: modelo completo (sem dados faltantes), método de deleção *listwise*, método de imputação pela média, método de imputação por *Random Forest*, método de imputação por *hot-deck* e método de imputação múltipla.

Realizamos as simulações de acordo com os seguintes passos:

Passo 1: Geramos x_{2i} da distribuição Normal com média 0 e variância 1 e, em seguida, x_{1i} da distribuição Weibull com parâmetro de escala $\exp(\gamma_0 + \gamma_1 x_{2i})$ e parâmetro de forma α_2 e y_i da distribuição Weibull com parâmetro de escala $\exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})$ e parâmetro de forma α_1, para $i = 1, ..., n$, sendo n, portanto, o tamanho da amostra final. Os verdadeiros valores destes parâmetros considerados são: $\alpha_1 = 2,\ \alpha_2 = 2,\ \beta_0 = 0,\ \beta_1 = 0.4,\ \beta_2 = -0.4,\ \gamma_0 = -0.5$ e $\gamma_1 = -0.5$;

Passo 2: De acordo com o mecanismo considerado e o valor de $p,\ 0 \leq p \leq 1$, geramos valores faltantes na variável X_1 da mesma forma que descrito na Seção 4.2.2. Ressaltamos que no caso de estimação pelo modelo completo, para comparação de desempenho, esse passo não é realizado;

Passo 3: Para cada observação i, que possui valor faltante em relação a X_1, geramos M valores para x_{1i} da distribuição de $X_1|X_2$, ou seja, da distribuição Weibull com parâmetro de forma dado, inicialmente, de maneira não informativa, como 1 e parâmetro de escala $\exp(\gamma_0 + \gamma_1 x_{2i})$, com $\gamma_0 = 0$ e $\gamma_1 = 0$ como valores iniciais. Para os métodos de integração numérica e média de log-verossimilhanças, em todos os cenários considerados neste estudo de simulação, foram geradas $M = 500$ valores para cada observação x_{1i} faltante. Entretanto,

como o tempo para processamento do método por algoritmo EM é relativamente maior, reduzimos o valor de M para esse método. Em alguns cenários do algoritmo EM (MCAR $n = 100$ e $p = 20\%$; MCAR $n = 300$ e $p = 20\%$; MCAR $n = 100$ e $p = 60\%$ para as primeiras 6 réplicas; MCAR $n = 300$ e $p = 60\%$ para as primeiras 3 réplicas) foram mantidos os $M = 500$ valores da distribuição de $X_1|X_2$, mas em todos os outros cenários e réplicas consideramos $M = 50$ valores. Para os cenários em que temos réplicas com $M = 500$ e $M = 50$, fizemos análises de sensibilidade dos resultados a esses dois valores e os resultados foram muito similares;

Passo 4: Maximizamos as funções de log-verossimilhança utilizando o *optim* do software estatístico R com parâmetro *fnscale* $= -1$. O método numérico usado para encontrar o máximo das funções de log-verossimilhança é o de Nelder-Mead, que é comumente aplicado em problemas de otimização não-linear para os quais as derivadas não podem ser encontradas ou não são definidas. Para obtermos melhores estimativas dos parâmetros, fazemos uma primeira maximização usando como valores iniciais valores não informativos e, em seguida, maximizamos as funções de log-verossimilhança novamente, também utlizando o método de Nelder-Mead e considerando, como valores iniciais, os valores das estimativas obtidos pelo primeiro processo de maximização. Quanto às funções de log-verossimilhança maximizadas, os procedimentos para os métodos de imputação pela média, imputação por *Random Forest*, imputação por *Hot-Deck*, imputação múltipla, deleção de casos e considerando o conjunto de dados completo, são as mesmas descritas na Seção 4.2.2, porém agora a função de log-verossimilhança a ser maximizada é:

$$\sum_{i=1}^{n} \log[f(y_i|x_{1i}, x_{2i}, \alpha_1, \beta_0, \beta_1, \beta_2)]$$

sendo f a densidade da distribuição Weibull com parâmetro de forma α_1 e parâmetro de escala $\exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})$. Para os métodos propostos e desenvolvidos neste trabalho, as funções de log-verossimilhança maximizadas são:

a) Para o método por integração numérica, utilizamos a função de log-verossimilhança construída na Seção 5.1.1, que considera os dois cenários possíveis para cada observação do conjunto de treinamento ou estimação, ou seja, a observação i ser observada em relação à variável X_1 ou a observação i ser faltante em X_1;

b) Para o método utilizando média de log-verossimilhanças, utilizamos a função de log-verossimilhança construída na Seção 5.1.2;

c) Para o método utilizando o algoritmo EM, consideramos a função de log-verossimilhança construída na Seção 5.1.3.

Passo 5: Para os três métodos sem resolução analítica aqui propostos, repetimos os passos 3 e 4 até a convergência, ou seja, até a diferença entre os valores estimados de cada parâmetro entre

uma iteração e a iteração seguinte ser menor do que 10^{-3}, podendo o número máximo de iterações para convergência ser igual a 10000;

Passo 6: Após obtermos os valores das estimativas dos parâmetros para cada conjunto de dados dentro de cada método, calculamos a diferença e a diferença quadrática dessas estimativas em relação aos verdadeiros valores dos parâmetros. Como os parâmetros γ_0, γ_1 e α_2 são estimados diretamente apenas pelas metodologias propostas, os índices de desempenho dos seus estimadores não são mostrados e comparados;

Passo 7: Para analisarmos os métodos em relação ao poder preditivo, calculamos o erro quadrático médio do $\widehat{y}_i$ em relação ao y_i observado para todas as observações do conjunto de teste. O cálculo de $\widehat{y}_i$ se dá da seguinte forma:

a) Para as metodologias propostas e cada observação i do conjunto de dados de teste, calculamos $\widehat{y}_i$ de acordo com o proposto na Seção 5.1.4. Vale ressaltar que, para cada caso em que a observação i é faltante em relação a X_1, geramos 500 (ou, em alguns casos para o método utilizando o algoritmo EM, geramos 50) valores de x_{1i} da distribuição de $X_1|X_2$, ou seja, da distribuição Weibull com parâmetro de forma dado por $\widehat{\alpha_2}$ (α_2 estimado) e parâmetro de escala $\exp(\widehat{\gamma_0} + \widehat{\gamma_1} x_{2i})$, sendo $\widehat{\gamma_0}$ e $\widehat{\gamma_1}$ as estimativas dos parâmetros γ_0 e γ_1 obtidas por cada método e em cada réplica;

b) Para os métodos de imputação de dados por *Random Forest*, imputação por *Hot-Deck* e imputação múltipla, realizamos a imputação dos dados faltantes na base de teste usando os mesmos procedimento descritos na Seção 4.2.2 e, com os dados completos, calculamos $\widehat{y}_i$ de acordo com o item i) da Sessão 5.1.4. Para o caso da imputação múltipla, como criamos cinco conjuntos de dados completos, o erro quadrático médio é dado pela média entre os erros quadráticos médios dos cinco conjuntos gerados;

c) Para o método de imputação pela média, a média das observações não faltantes em X_1 do conjunto de treino é o valor utilizado para ser imputado nas observações que possuem valores faltantes no conjunto de teste e, então, após esta imputação, calculamos $\widehat{y}_i$ de acordo com o item i) da Sessão 5.1.4;

d) Para o método em que consideramos o conjunto de dados de teste completo, sem valores faltantes, calculamos $\widehat{y}_i$ de acordo com o item i) da Sessão 5.1.4;

e) Para o método de deleção de dados, como deletamos as observações que possuem valores faltantes, não conseguimos calcular o $\widehat{y}_i$ associado a elas pois não existe, de fato, um processo de imputação ou predição de valores faltantes. Logo, não analisamos o desempenho de predição desse método nas observações do conjunto de teste, por não fazer sentido essa comparação.

As Figuras 49, 51, 53 e 55 mostram o desempenho inferencial dos métodos comparados para os 4 cenários analisados com mecanismo MCAR. Em vez de calcularmos um único valor

de viés e EQM para a estimativa de cada parâmetro, como sendo a média das diferenças (viés) e das diferenças quadráticas (EQM) observadas entre estimativa e parâmetro nas 30 réplicas, preferimos exibir todas diferenças observadas através de boxplots, para melhor comparação. As Figuras 50, 52, 54 e 56 mostram os erros quadráticos médios de predição observados para as mesmas simulações.

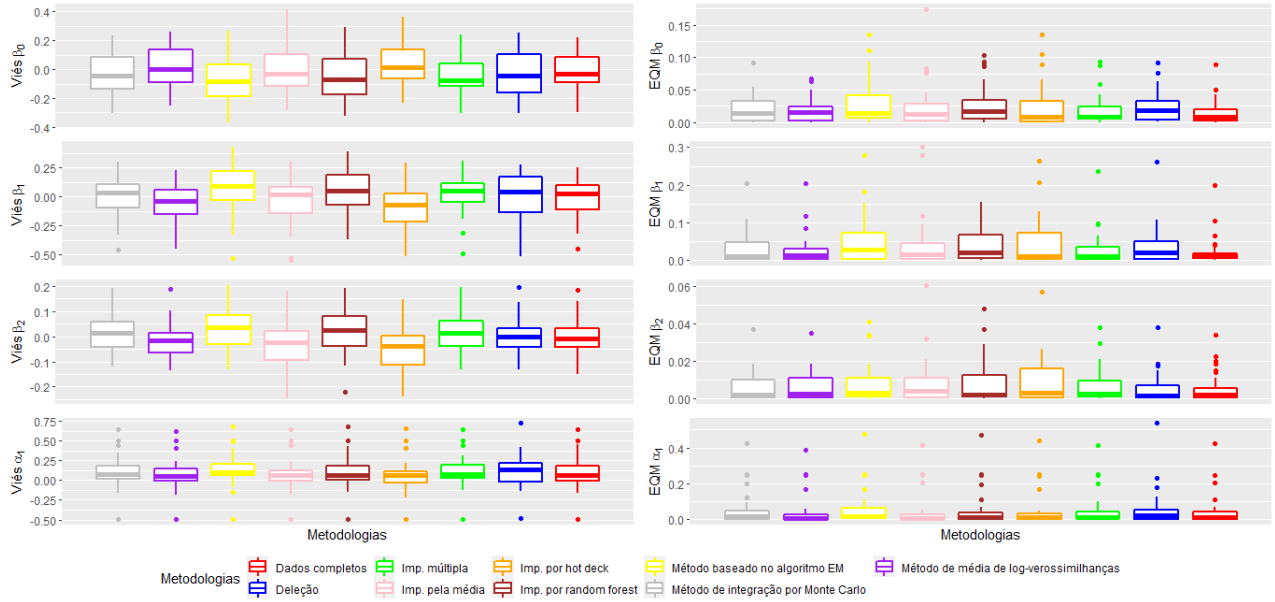

Figura 49 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário MCAR, $n = 100$ e $p = 0.20$.

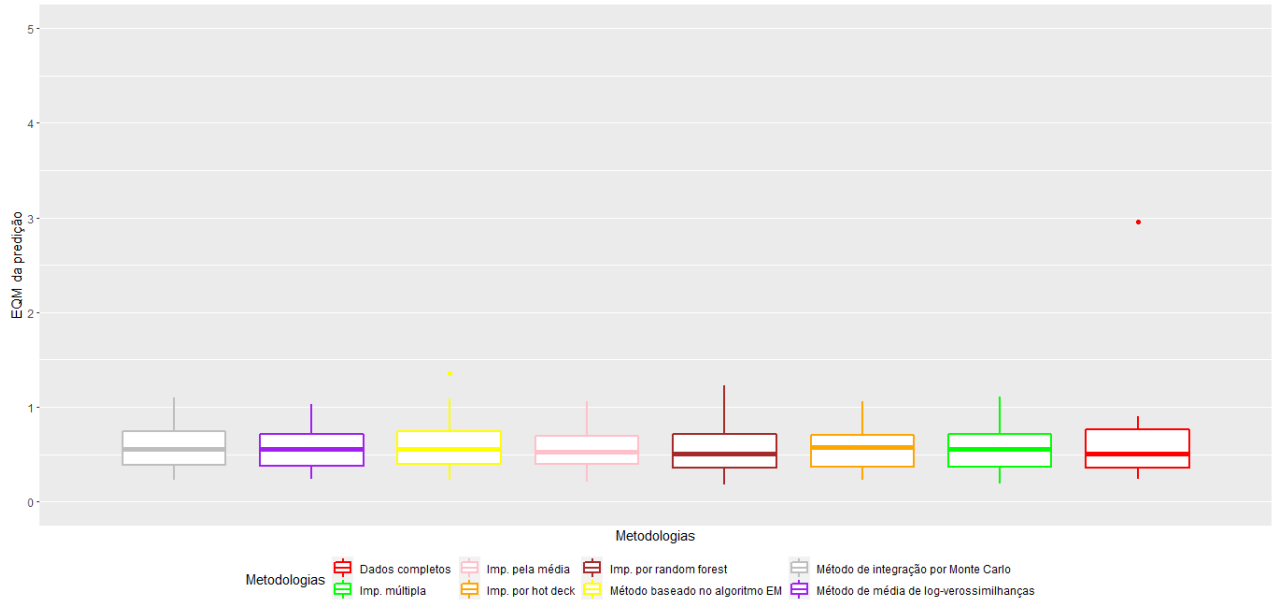

Figura 50 – Erro quadrático médio dos valores preditos para o cenário MCAR, $n = 100$ e $p = 0.20$.

Para este cenário MCAR, $n = 100$ e $p = 0.20$, foram detectados mais *outliers*, além dos já exibidos nas figuras. Eles não constam no gráfico, pois reduziriam consideralmente a escala dos boxplots e dificultaria, com isso, a análise dos desempenhos dos métodos. Seguem os valores dos *outliers* detectados que não constam nos gráficos:

- Método de integração por Monte Carlo: 7.40;
- Método de imputação pela média: 5.85;
- Método de imputação por *random forest:* 6.12;
- Método de imputação por *hot-deck*: 5.33;
- Método de imputação múltipla: 7.14;
- Dados completos: 5.94;
- Método baseado no algoritmo EM: 6.53;
- Método de média de log-verossimilhanças: 7.78.

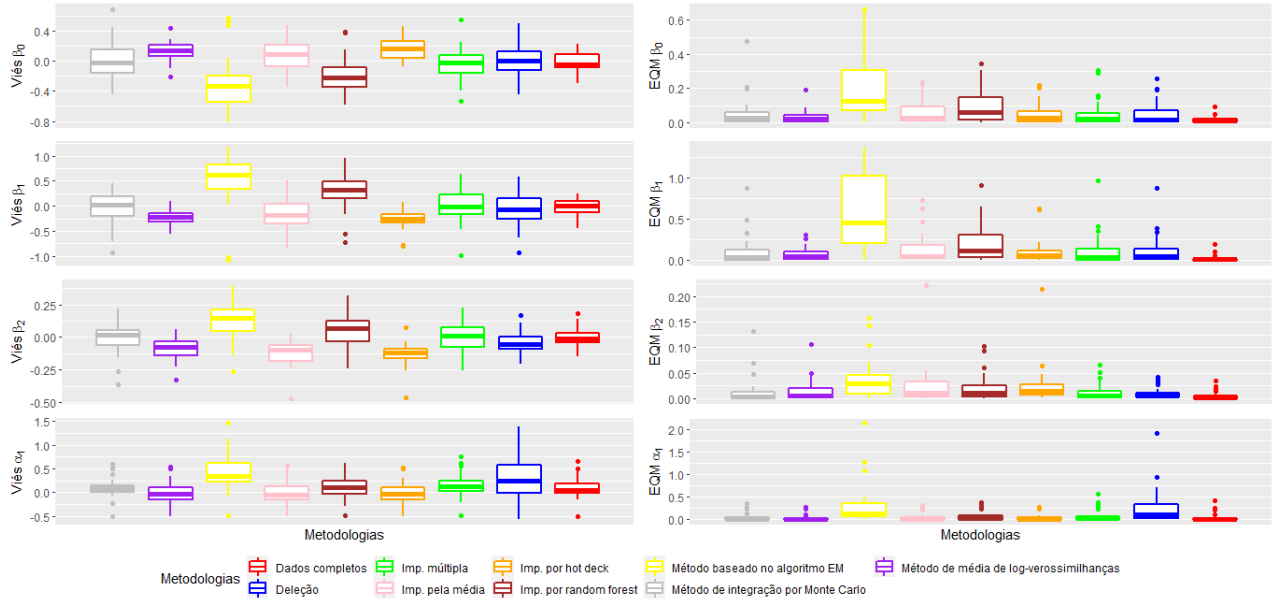

Figura 51 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário MCAR, $n = 100$ e $p = 0.60$.

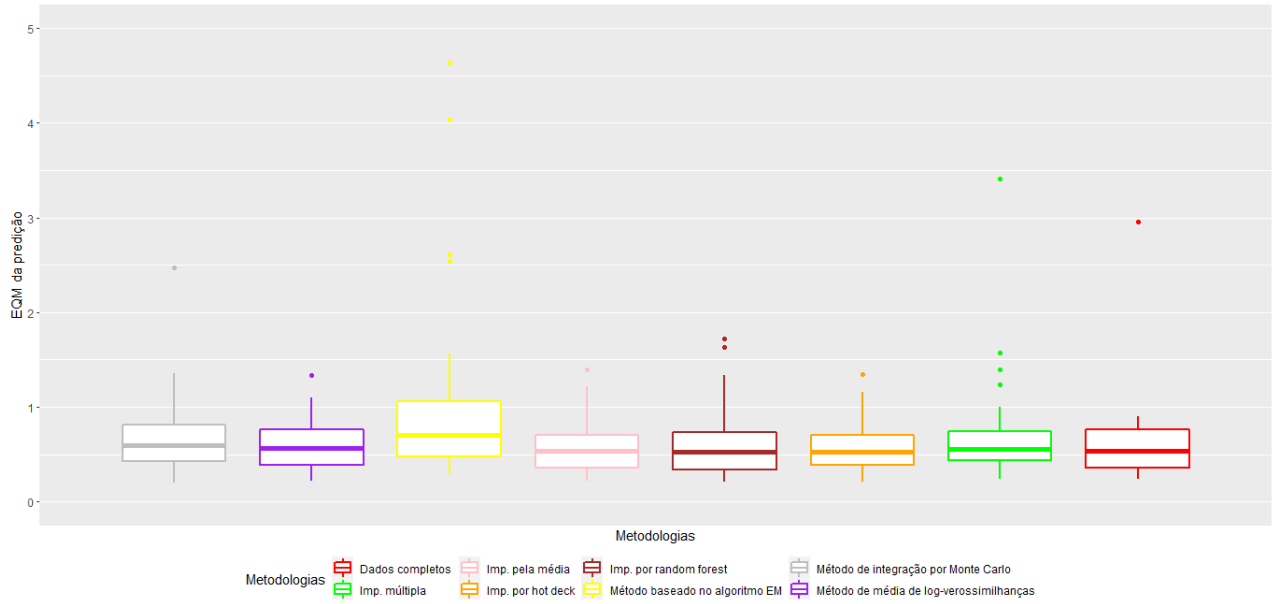

Figura 52 – Erro quadrático médio dos valores preditos para o cenário MCAR, $n = 100$ e $p = 0.60$.

Os *outliers* detectados que não constam nos gráficos para o cenário MCAR, $n = 100$ e $p = 0.60$, são:

- Método de integração por Monte Carlo: 10.26;
- Método de imputação pela média: 10.88;
- Método de imputação por *random forest:* 7.97;
- Método de imputação por *hot-deck*: 10.51;
- Método de imputação múltipla: 36.20;
- Dados completos: 5.94;
- Método baseado no algoritmo EM: 5.01, 6.26, 9.26;
- Método de média de log-verossimilhanças: 10.48.

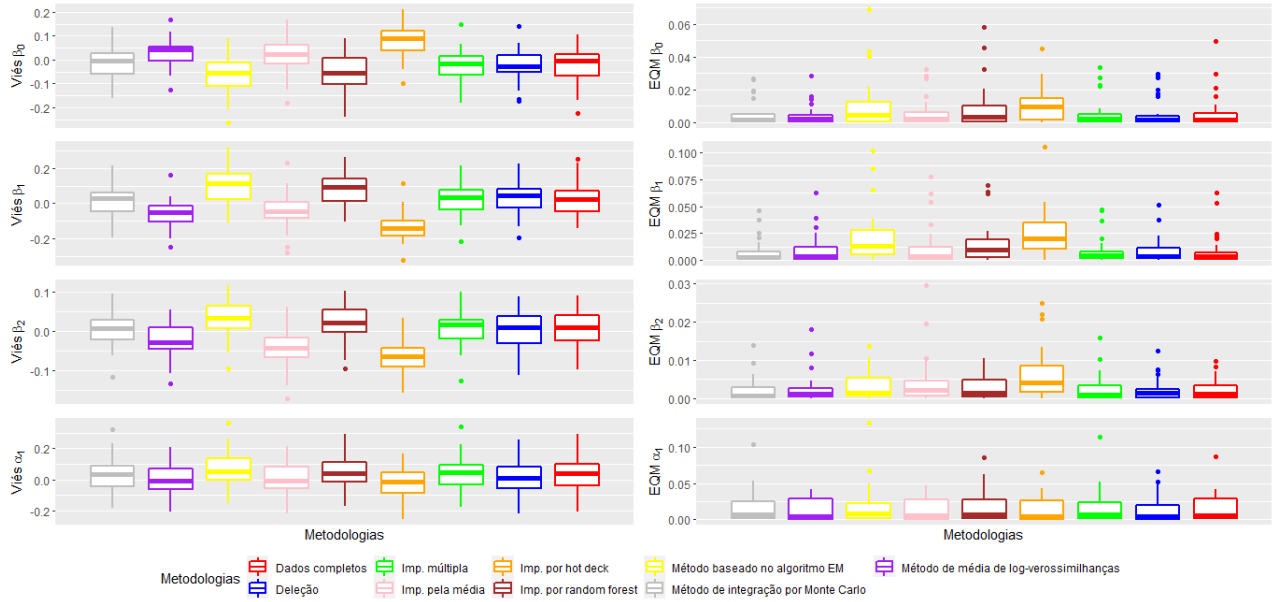

Figura 53 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário MCAR, $n = 300$ e $p = 0.20$.

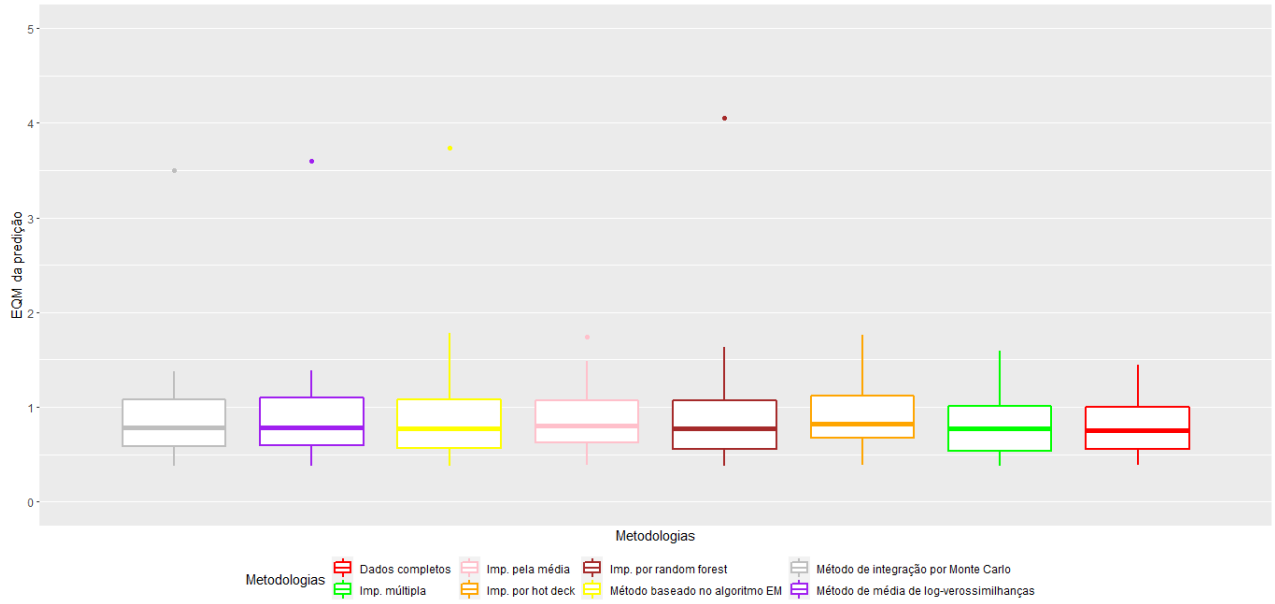

Figura 54 – Erro quadrático médio dos valores preditos para o cenário MCAR, $n = 300$ e $p = 0.20$.

Seguem os *outliers* detectados que não constam nos gráficos para o cenário MCAR, $n = 300$ e $p = 0.20$:

- Método de imputação pela média: 5.70;
- Método de imputação por *hot-deck*: 5.09;
- Método de imputação múltipla: 13.04;
- Dados completos: 9.18, 92.92.

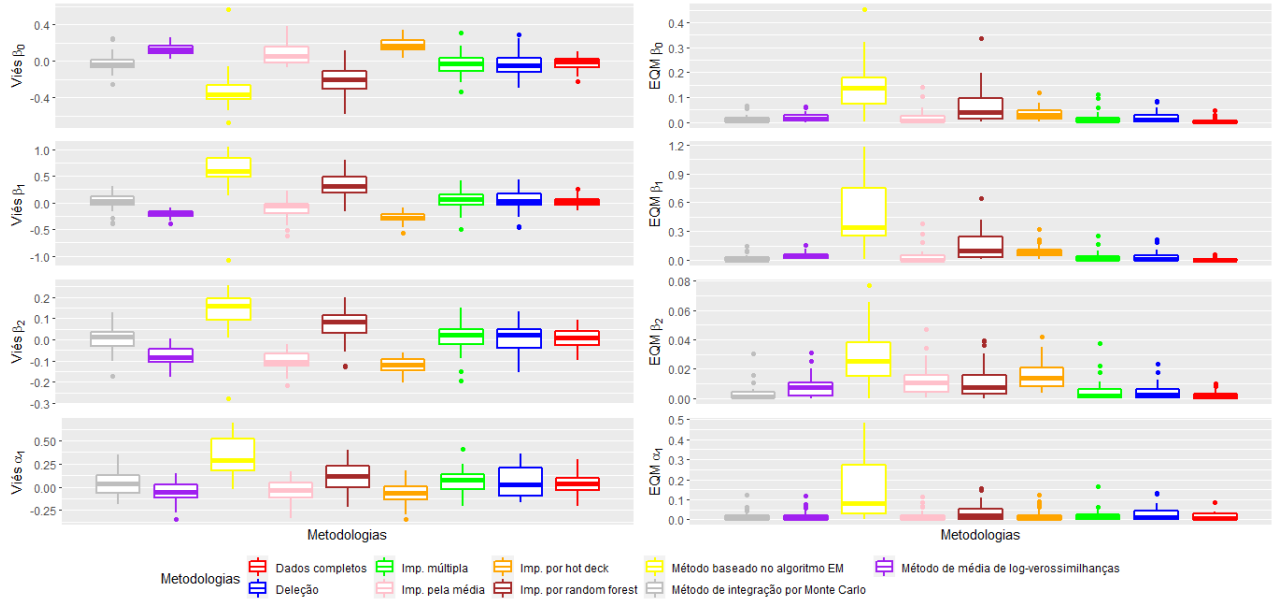

Figura 55 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário MCAR, $n = 300$ e $p = 0.60$.

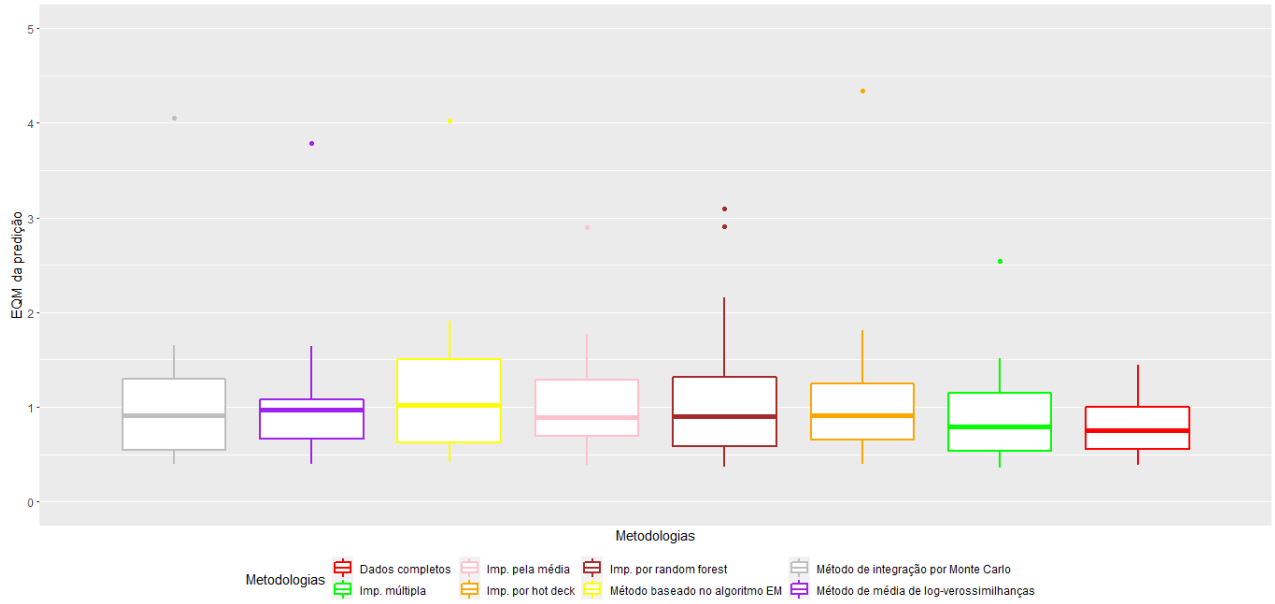

Figura 56 – Erro quadrático médio dos valores preditos para o cenário MCAR, $n = 300$ e $p = 0.60$.

Os *outliers* detectados que não constam nos gráficos para o cenário MCAR, $n = 300$ e $p = 0.60$, são:

- Método de imputação pela média: 5.07;
- Método de imputação por *random forest:* 6.10, 14.41;
- Método de imputação múltipla: 7.80, 22.32, 452.78;
- Dados completos: 9.18, 92.92;
- Método baseado no algoritmo EM: 6.25, 7.65, 8.65, 8.74, 9.68, 13.07, 36.97, 57.40.

Considerando o mecanismo MCAR de geração de dados faltantes, observamos que a metodologia proposta utilizando integração por Monte Carlo apresenta resultados inferenciais muito parecidos à análise dos dados completos, especialmente para amostras maiores ($n = 300$). Esse método se sobressai inclusive ao método de imputação múltipla, que geralmente apresenta boa performance no mecanismo MCAR. Isso representa uma ótima performance do método na estimação de parâmetros na presença de dados faltantes. Por outro lado, o método utilizando o algoritmo EM e os métodos por imputação por *Random Forest*, *Hot-deck*, e por média apresentam performances inferiores.

O desempenho ruim do método utilizando o algoritmo EM se deve ao fato desse algoritmo ser sensível ao ponto inicial. Como estamos considerando valores não informativos para os parâmetros no início dos métodos, o método pode talvez não convergir para o máximo global e, em algumas réplicas, não convergir.

Vale observar também que, para as amostras menores ($n = 100$), apesar do viés das estimativas do método de média de log-verossimilhanças ser pior que o do método da integração por Monte Carlo, ele apresenta menores diferenças quadráticas. Isso talvez se deva ao fato de que a variância dos estimadores por esse método seja menor que a variância dos estimadores por integração de Monte Carlo e isso compensou o viés no cálculo do EQM.

Em relação ao desempenho preditivo, exceto em alguns detalhes ou *outliers*, os métodos apresentaram comportamento muito parecido. Esse comportamento faz sentido, pois o mecanismo de geração dos dados faltantes é o completamente aleatório, quando é esperado que todas as metodologias performem de maneira razoável. Para amostras menores, o método utilizando média de log-verossimilhanças obtém melhores resultados quanto ao desempenho preditivo em relação aos nossos métodos propostos. Já para as amostras maiores, o método por integração por Monte Carlo se sobressai.

As Figuras 57, 59, 61 e 63 apresentam os resultados inferenciais dos métodos para o mecanismo MAR de dados faltantes e as Figuras 58, 60, 62 e 64 o desempenho preditivo.

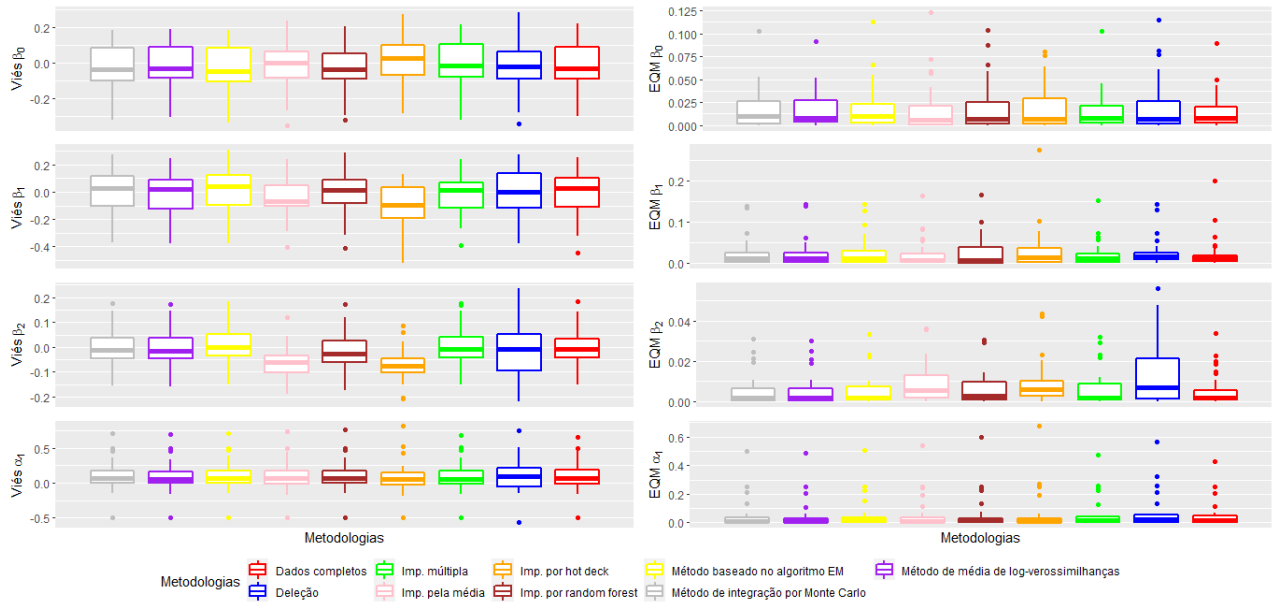

Figura 57 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário MAR, $n = 100$ e $p = 0.20$.

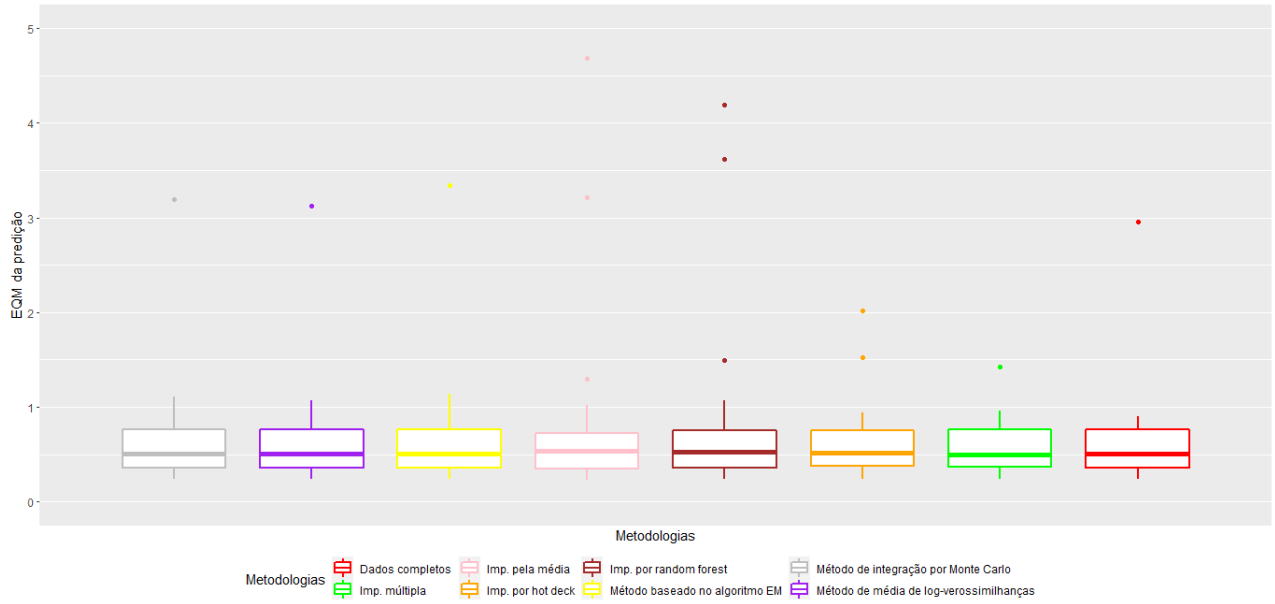

Figura 58 – Erro quadrático médio dos valores preditos para o cenário MAR, $n = 100$ e $p = 0.20$.

Seguem os *outliers* detectados que não constam nos gráficos no cenário MAR, $n = 100$ e $p = 0.20$:

- Método de integração por Monte Carlo: 5.41;
- Método de imputação por *hot-deck*: 8.81;
- Método de imputação múltipla: 5.60;
- Dados completos: 5.94;
- Método baseado no algoritmo EM: 5.21;
- Método de média de log-verossimilhanças: 5.55.

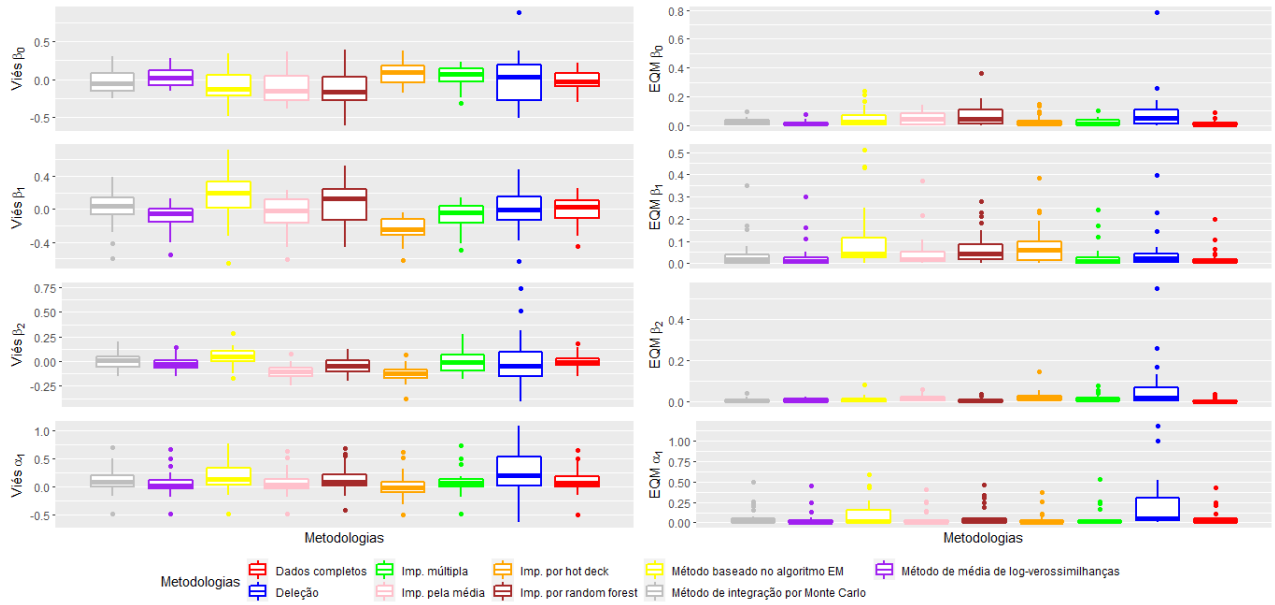

Figura 59 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário MAR, $n = 100$ e $p = 0.60$.

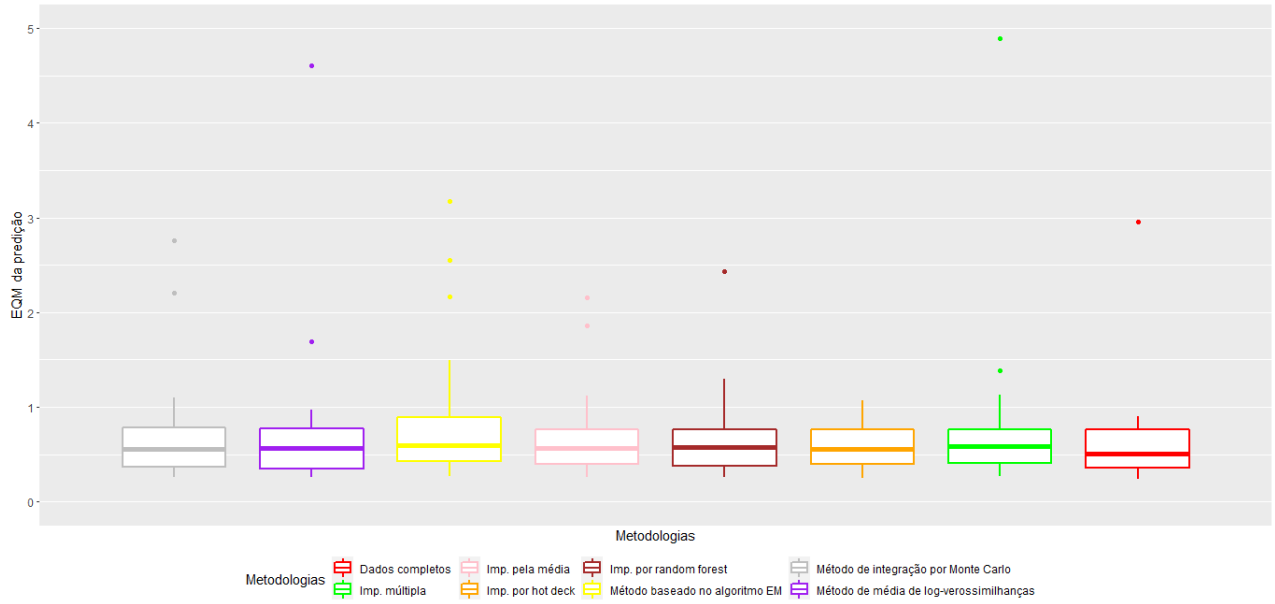

Figura 60 – Erro quadrático médio dos valores preditos para o cenário MAR, $n = 100$ e $p = 0.60$.

Os *outliers* detectados que não constam nos gráficos no cenário MAR, $n = 100$ e $p = 0.60$, são:

- Método de imputação por *random forest:* 146.50;
- Método de imputação por *hot-deck*: 7.75;
- Dados completos: 5.94;

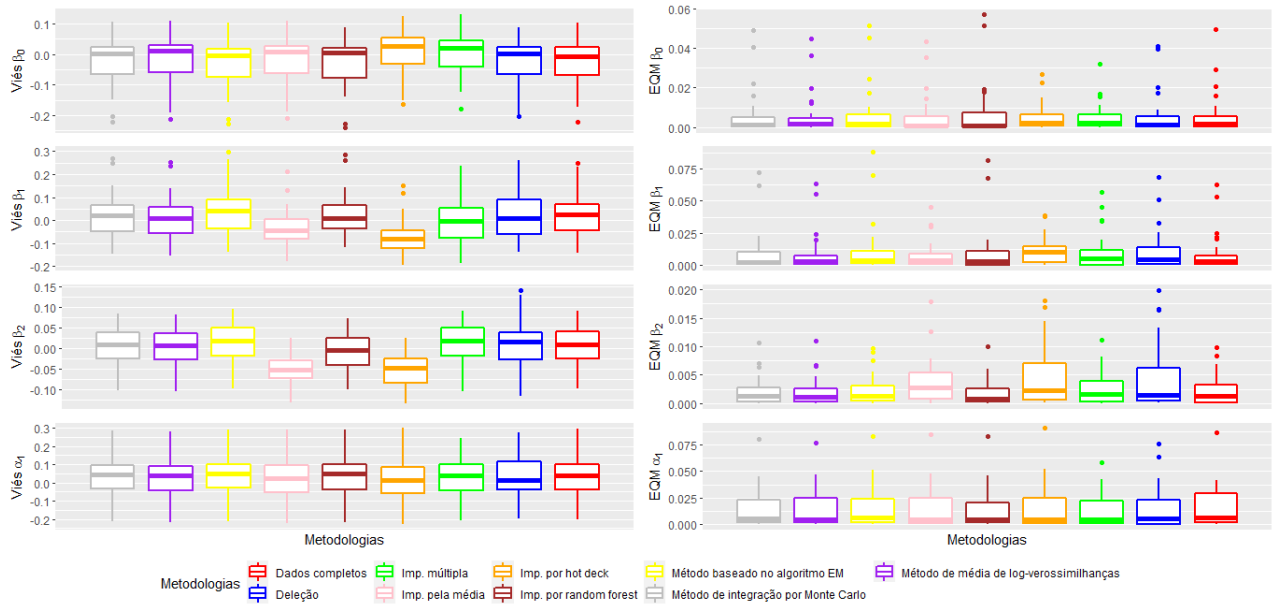

Figura 61 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário MAR, $n = 300$ e $p = 0.20$.

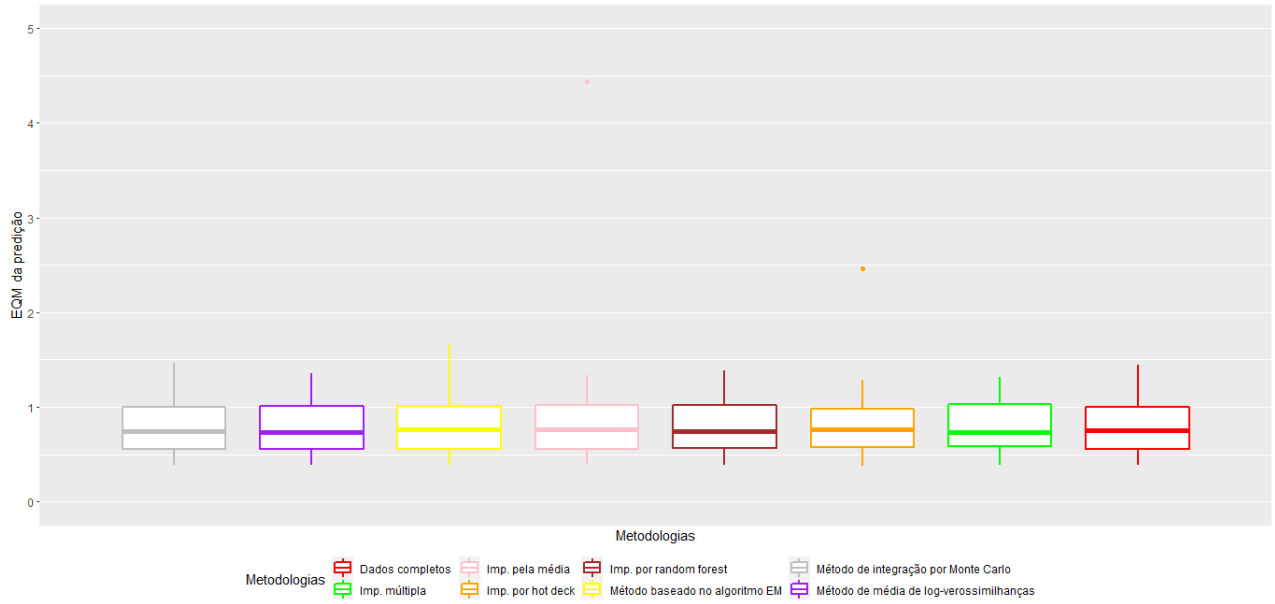

Figura 62 – Erro quadrático médio dos valores preditos para o cenário MAR, $n = 300$ e $p = 0.20$.

Seguem os *outliers* detectados que não constam nos gráficos no caso MAR, $n = 300$ e $p = 0.20$:

- Método de integração por Monte Carlo: 8.09, 161.62;
- Método de imputação pela média: 116.32;
- Método de imputação por *random forest:* 6.32, 228.18;
- Método de imputação por *hot-deck*: 43.30;
- Método de imputação múltipla: 8.42, 113.60;
- Dados completos: 9.18, 92.92;
- Método baseado no algoritmo EM: 9.00, 219.76;
- Método de média de log-verossimilhanças: 7.06, 129.90.

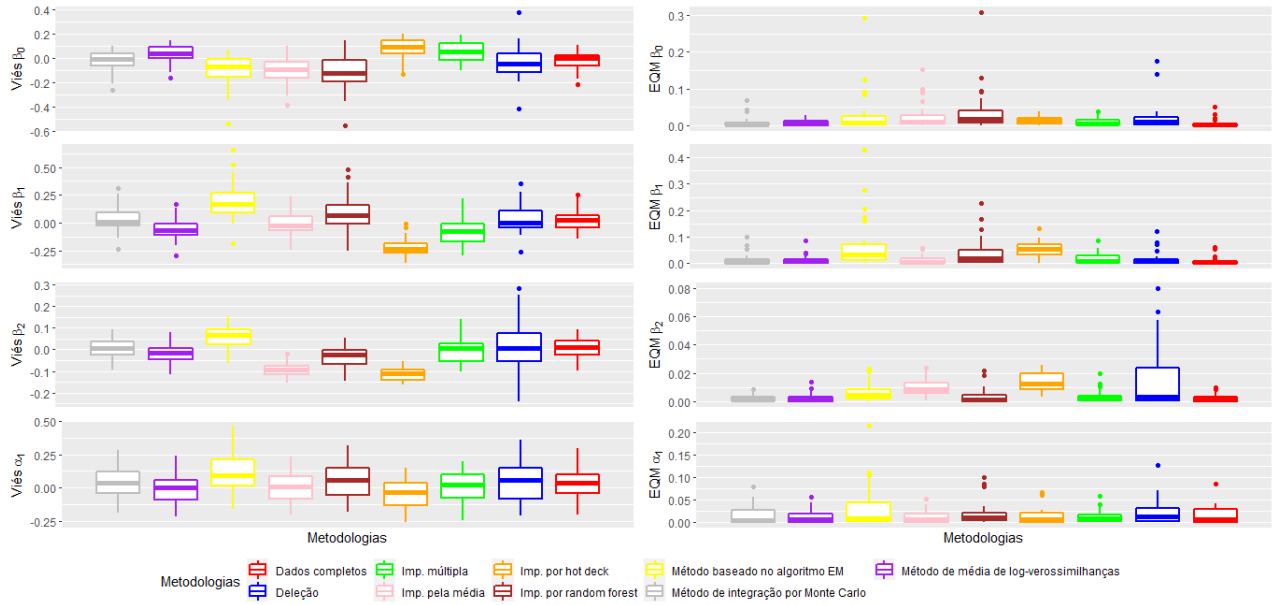

Figura 63 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário MAR, $n = 300$ e $p = 0.60$.

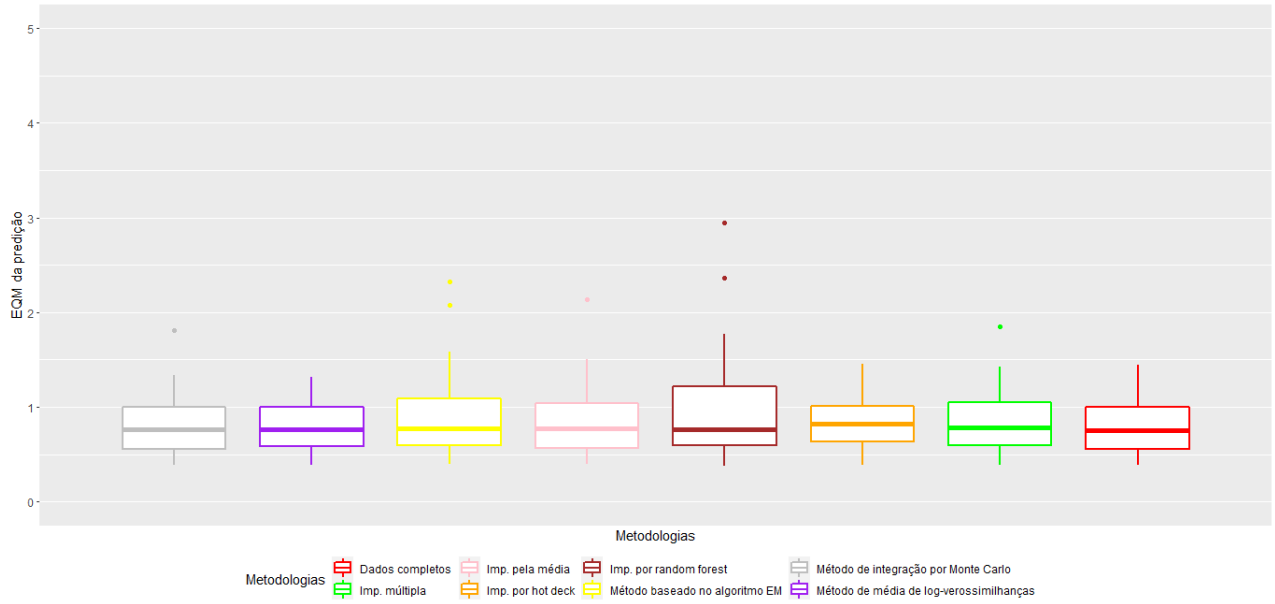

Figura 64 – Erro quadrático médio dos valores preditos para o cenário MAR, $n = 300$ e $p = 0.60$.

Os *outliers* detectados que não constam nos gráficos do cenário MAR, $n = 300$ e $p = 0.60$, são:

- Método de integração por Monte Carlo: 26.57, 266.72;
- Método de imputação pela média: 42.06, 180.67;
- Método de imputação por *random forest:* 22.03, 627.06;
- Método de imputação por *hot-deck*: 6.59;
- Método de imputação múltipla: 94.90;
- Dados completos: 9.18, 92.92;
- Método baseado no algoritmo EM: 8.35, 93.30, 2706.76;
- Método de média de log-verossimilhanças: 10.86, 42.58.

Para o mecanismo de geração dos dados faltantes MAR, observamos que o método utilizando média de log-verossimilhanças apresenta resultados inferenciais muito parecidos com a estimação considerando os dados completos e o método de imputação múltipla. Para os cenários com maior proporção de valores faltantes ($p = 60\%$) ele se sobressai em relação ao método de imputação múltipla, tanto em relação ao viés e erro quadrático médio das estimativas dos parâmetros, quanto em relação à performance preditiva.

O método utilizando integração por Monte Carlo também apresenta resultados inferenciais e preditivos muito satisfatórios, especialmente em relação ao viés das estimativas dos parâmetros, em que ele se equipara ao método de imputação múltipla e, para amostras maiores ($n = 300$), consegue se sobressair a ele.

Vale observarmos que, para este cenário MAR, torna-se mais evidente que o método baseado na delação de dados apresenta um erro quadrático médio das estimativas dos parâmetros maior quando comparado aos demais e isso faz sentido porque o tamanho da amostra de estimação fica bem menor quando comparado ao tamanho da amostra de estimação dos outros métodos e isso promove uma alta variabilidade nas estimativas dos parâmetros, apesar do viés não ser tão grande.

Apesar do método de estimação por algoritmo EM e métodos de imputação por média, *random forest* e *hot-deck* apresentarem um desempenho preditivo relativamente similar aos demais, seus desempenhos inferenciais são geralmente inferiores.

As Figuras 65, 67, 73 e 71 apresentam os resultados inferenciais dos métodos para o mecanismo MNAR de dados faltantes e as Figuras 66, 68, 74 e 72 o desempenho preditivo.

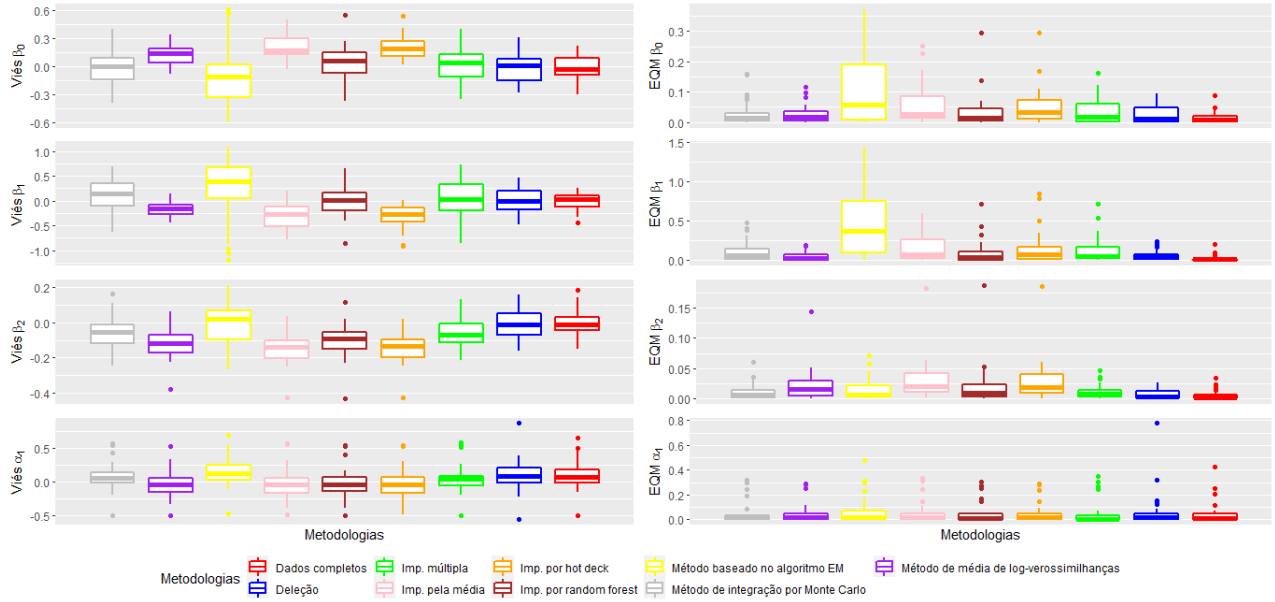

Figura 65 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário MNAR, $n = 100$ e $p = 0.20$.

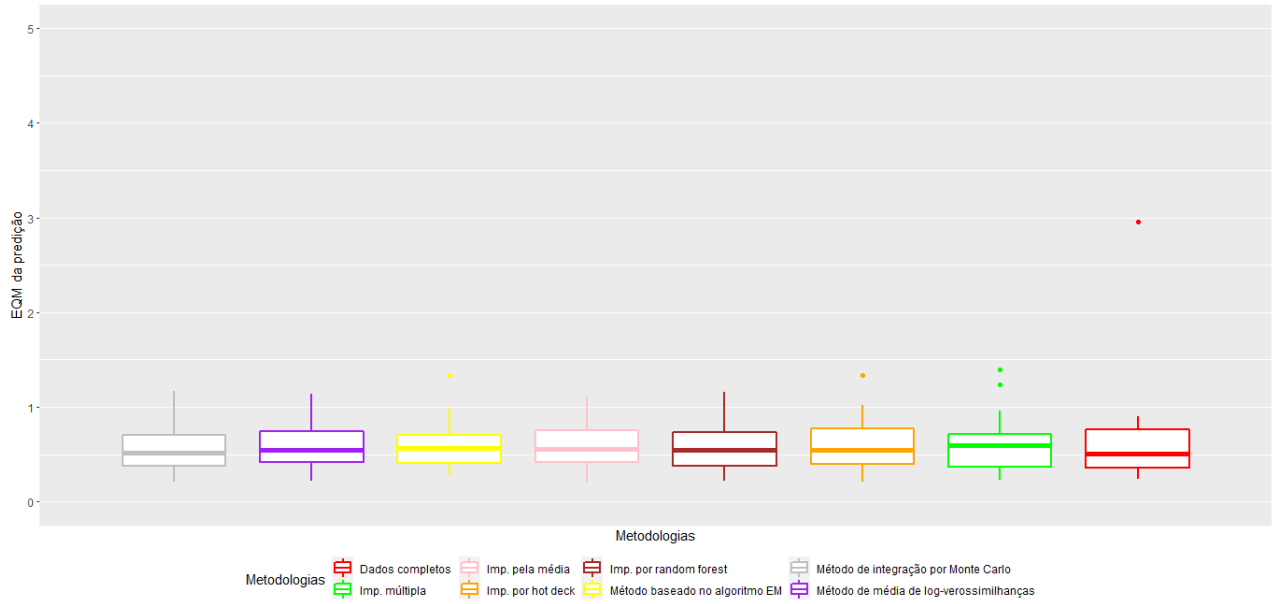

Figura 66 – Erro quadrático médio dos valores preditos para o cenário MNAR, $n = 100$ e $p = 0.20$.

Valores *outliers* detectados que não constam nos gráficos para o cenário MNAR, $n = 100$ e $p = 0.20$, são:

- Método de integração por Monte Carlo: 11.13;
- Método de imputação pela média: 10.6;
- Método de imputação por *random forest:* 11.06;
- Método de imputação por *hot-deck*: 10.51;
- Método de imputação múltipla: 11.04;
- Dados completos: 5.94;
- Método baseado no algoritmo EM: 12.11;
- Método de média de log-verossimilhanças: 11.04.

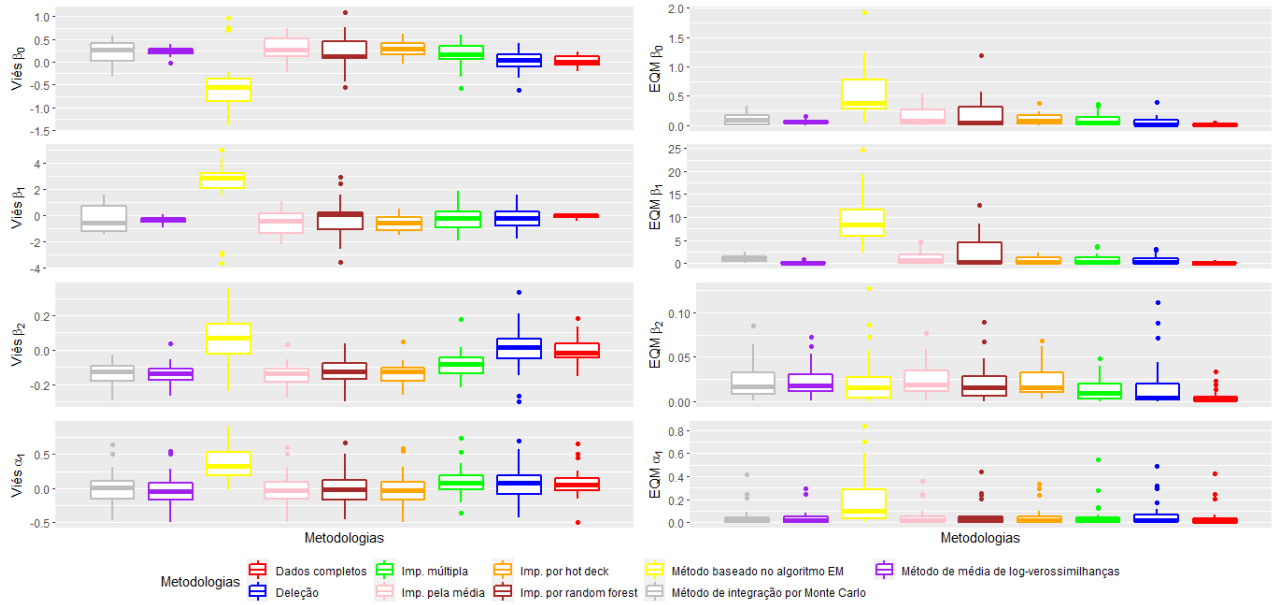

Figura 67 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário MNAR, $n = 100$ e $p = 0.60$.

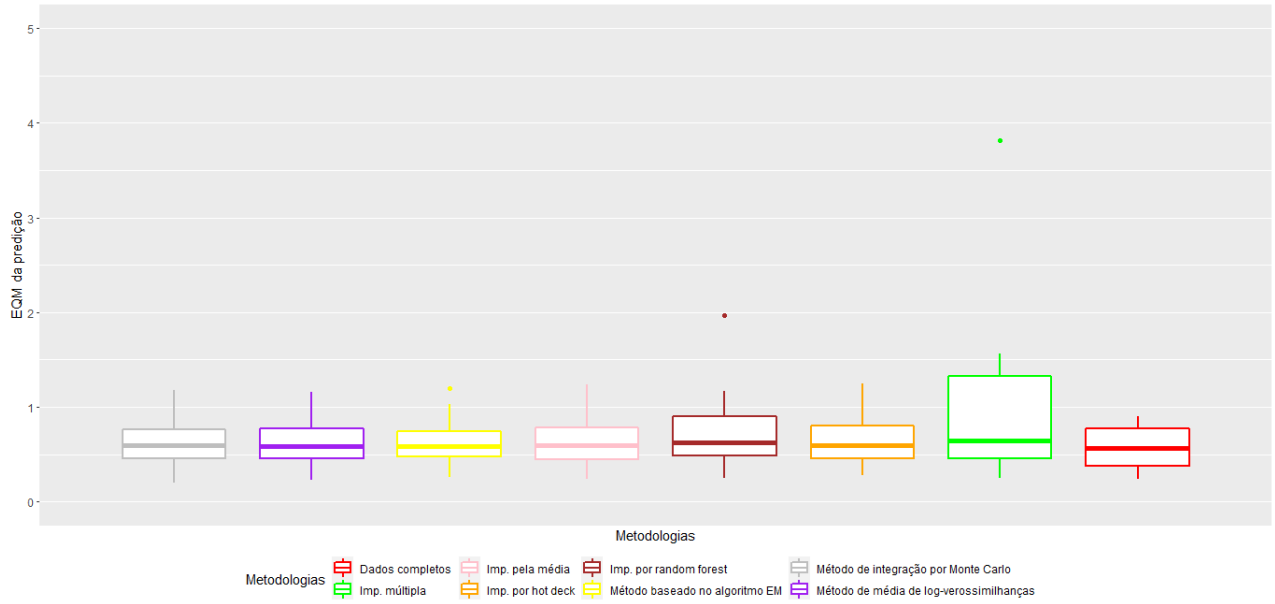

Figura 68 – Erro quadrático médio dos valores preditos para o cenário MNAR, $n = 100$ e $p = 0.60$.

Seguem os *outliers* detectados que não constam nos gráficos no caso MNAR, $n = 100$ e $p = 0.60$:

- Método de integração por Monte Carlo: 10.84;
- Método de imputação pela média: 11.09;
- Método de imputação por *random forest:* 12.59;
- Método de imputação por *hot-deck*: 10.70;
- Método de imputação múltipla: 8.39, 749.91;
- Dados completos: 5.94;
- Método baseado no algoritmo EM: 12.14;
- Método de média de log-verossimilhanças: 10.82.

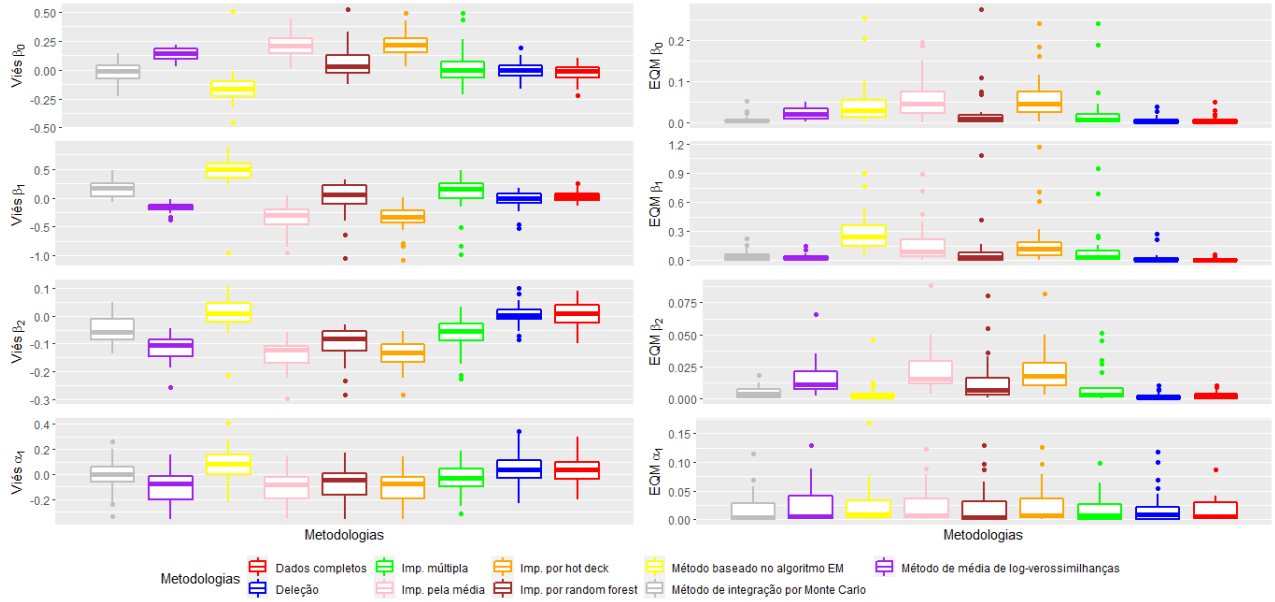

Figura 69 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário MNAR, $n = 300$ e $p = 0.20$.

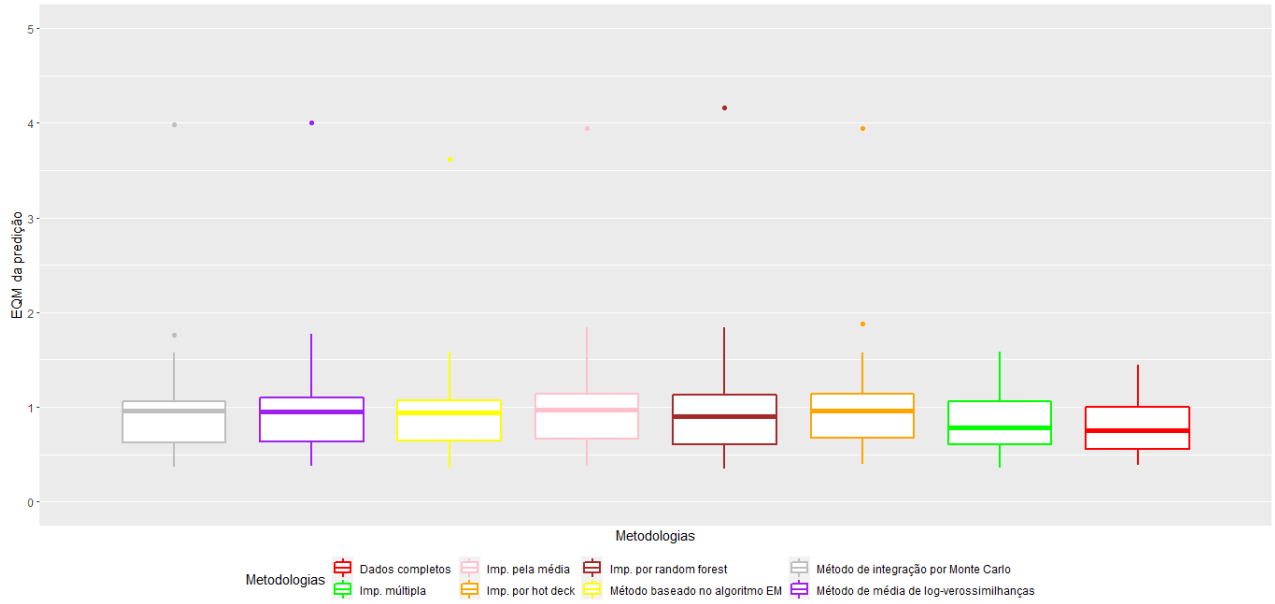

Figura 70 – Erro quadrático médio dos valores preditos para o cenário MNAR, $n = 300$ e $p = 0.20$.

Os *outliers* detectados que não constam nos gráficos para o cenário MNAR, $n = 300$ e $p = 0.20$, são:

- Método de imputação múltipla: 24.19;
- Dados completos: 9.18, 92.92;

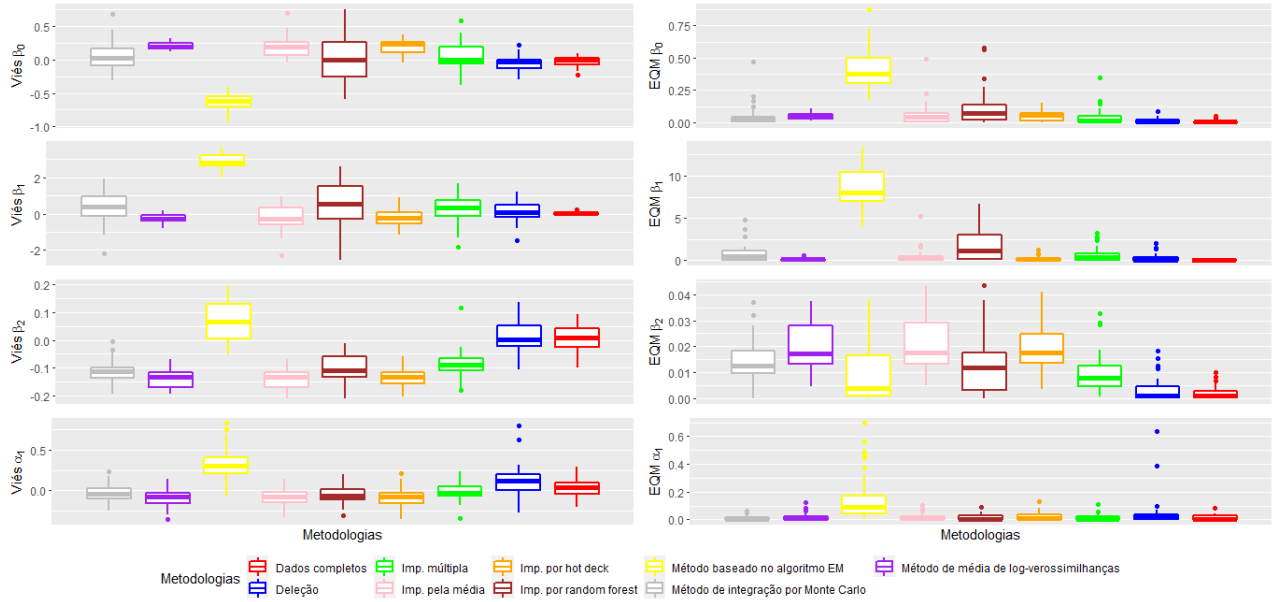

Figura 71 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário MNAR, $n = 300$ e $p = 0.60$.

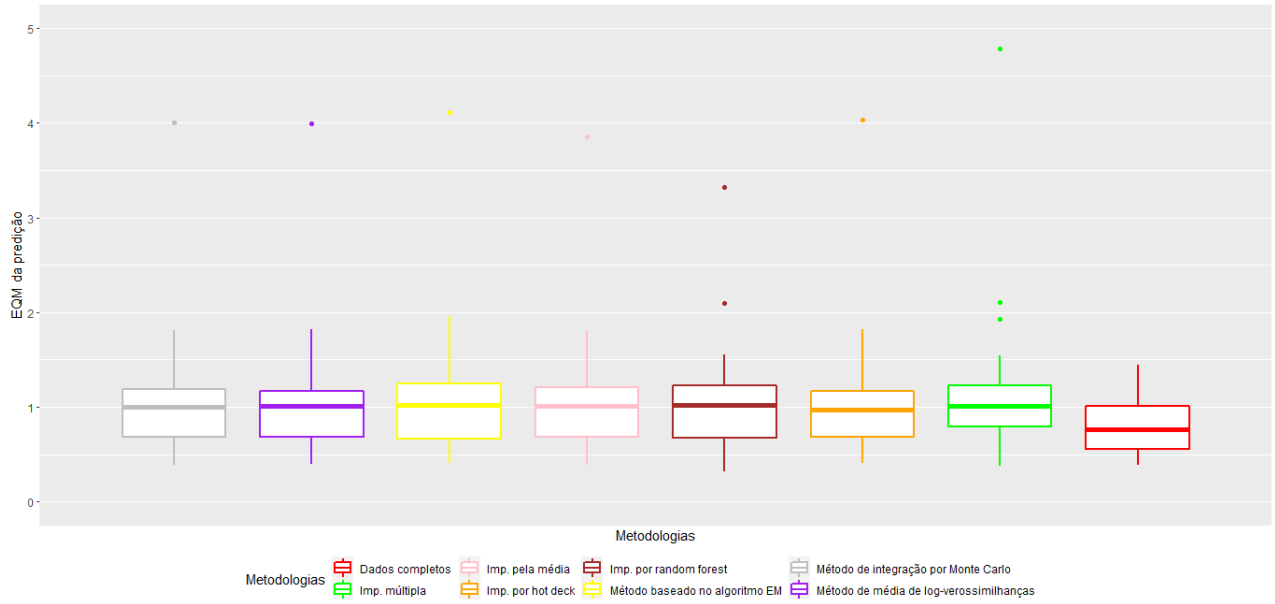

Figura 72 – Erro quadrático médio dos valores preditos para o cenário MNAR, $n = 300$ e $p = 0.60$.

Outros valores *outliers* no cenário MNAR, $n = 300$ e $p = 0.60$, foram observados apenas para o modelo estimado com dados completos e são: 9.18 e 92.92.

Considerando o mecanismo MNAR de geração dos dados faltantes, que provavelmente é a situação de estimação e predição mais desafiadora, de maneira geral, os métodos baseados em modelo utilizando integração por Monte Carlo e média de log-verossimilhanças e os métodos de imputação múltipla, deleção de casos e dados completos foram os que apresentaram melhores resultados em relação à estimação dos parâmetros. Para amostras menores ($n = 100$), levando-se em consideração o erro quadrático médio das estimativas dos parâmetros, o método utilizando média de log-verossimilhanças e o com dados completos se destacam em relação a todos os outros métodos, apesar do método por integração por Monte Carlo apresentar viés menores. Para amostras maiores ($n = 300$), o método utilizando integração por Monte Carlo também apresenta ótima performance em relação a esta métrica.

Quanto ao poder preditivo, os três métodos baseados em modelo sem resolução analítica, para amostras menores, apresentam desempenho melhor do que metodologias que são focadas na predição, como o método de imputação múltipla.

5.2 Modelo para duas variáveis com valores faltantes

Sejam Y, X_1, X_2, X_3 variáveis aleatórias, ou seja, $k = 3$. Vamos considerar que X_1 e X_2 possuem valores faltantes. Logo, como vimos na Seção 4.1, dada uma amostra de tamanho n de cada uma das variáveis Y, X_1, X_2, X_3, $(y_1, ..., y_n, x_{11}, ..., x_{1n}, x_{21}, ..., x_{2n}, x_{31}, ..., x_{3n})$ temos quatro cenários possíveis para esse caso: a) a observação i tem todas as suas variáveis observadas; b) a observação i tem valor faltante em X_1; c) a observação i tem valor faltante em X_2; d) a observação i tem valores faltantes em X_1 e X_2. Nessa seção, apresentamos três métodos para estimação de modelos na presença desses 4 cenários para modelos sem resolução analítica. São eles: método por integração numérica, método utilizando média de verossimilhanças e método utilizando o algoritmo EM.

5.2.1 Método por integração numérica

Considerando primeiramente o método resolvido por integração numérica de Monte Carlos, temos:

a) A observação i tem todas as suas variáveis observadas.
Seja n_1 $(n_1 < n)$ o número de observações que não possuem valores faltantes. Neste caso, consideraremos a distribuição de $Y, X_1, X_2 | X_3$ para estimação dos parâmetros, ou seja:

$$L(\theta|\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) = \prod_{i=1}^{n_1} f(y_i|x_{1i}, x_{2i}, x_{3i}, \theta) f(x_{1i}|x_{2i}, x_{3i}, \theta) f(x_{2i}|x_{3i}, \theta); \qquad (5.24)$$

b) A observação i tem valor faltante em X_1.
Seja n_2 $(n_2 < n)$ o número de observações que possuem valor faltante apenas em X_1. Neste caso, queremos a distribuição $f(y_i|x_{2i},x_{3i},\theta)$, que é dada por:

$$\begin{aligned} f(y_i|x_{2i},x_{3i},\theta) &= \int f(y_i,x_{1i}|x_{2i},x_{3i},\theta)dx_{1i} \\ &= \int f(y_i|x_{1i},x_{2i},x_{3i},\theta)f(x_{1i}|x_{2i},x_{3i},\theta)dx_{1i} \\ &= E_{X_1|X_2,X_3}[f(y_i|X_{1i},x_{2i},x_{3i},\theta))]. \end{aligned} \tag{5.25}$$

Então, geramos M_1 valores aleatórios para esta observação, $x_{1i1},\ldots,x_{1iM_1}$, da distribuição de $X_1|X_2,X_3$, para $i = 1,2,\ldots,n_2$, e a aproximação Monte Carlo de (5.25), para esta densidade, se torna:

$$\begin{aligned} f(y_i|x_{2i},x_{3i},\theta) &= E_{X_1|X_2,X_3}[f(y_i|X_{1i},x_{2i},x_{3i},\theta)] \\ &\approx \frac{1}{M_1}\sum_{m_1=1}^{M_1} f(y_i|x_{1im_1},x_{2i},x_{3i},\theta). \end{aligned} \tag{5.26}$$

Logo, $L(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3)$, neste cenário, é:

$$\begin{aligned} L(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) &= \prod_{i=1}^{n_2} f(y_i|x_{2i},x_{3i},\theta) \\ &\approx \prod_{i=1}^{n_2}\left[\frac{1}{M_1}\sum_{m_1=1}^{M_1} f(y_i|x_{1im_1},x_{2i},x_{3i},\theta)\right]; \end{aligned} \tag{5.27}$$

c) A observação i tem valor faltante em X_2.
Seja n_3 $(n_3 < n)$ o número de observações que possuem valor faltante apenas em X_2. Neste caso, queremos a distribuição $f(y_i|x_{1i},x_{3i},\theta)$, que é dada por:

$$\begin{aligned} f(y_i|x_{1i},x_{3i},\theta) &= \int f(y_i,x_{2i}|x_{1i},x_{3i},\theta)dx_{2i} \\ &= \int f(y_i|x_{1i},x_{2i},x_{3i},\theta)f(x_{2i}|x_{1i},x_{3i},\theta)dx_{2i} \\ &= E_{X_2|X_1,X_3}[f(y_i|x_{1i},X_{2i},x_{3i},\theta))]. \end{aligned} \tag{5.28}$$

Então, geramos M_2 valores aleatórios para esta observação, $x_{2i1},\ldots,x_{2iM_2}$, da distribuição de $X_2|X_1,X_3$, para $i = 1,2,\ldots,n_3$, e a aproximação Monte Carlo de (5.28), para esta densidade, se torna:

$$\begin{aligned} f(y_i|x_{1i},x_{3i},\theta) &= E_{X_2|X_1,X_3}[f(y_i|x_{1i},X_{2i},x_{3i},\theta)] \\ &= \frac{1}{M_2}\sum_{m_2=1}^{M_2} f(y_i|x_{1i},x_{2im_2},x_{3i},\theta). \end{aligned} \tag{5.29}$$

Logo, $L(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3)$, neste cenário, é:

$$\begin{aligned} L(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) &= \prod_{i=1}^{n_3} f(y_i|x_{1i},x_{3i},\theta) \\ &\approx \prod_{i=1}^{n_3}\left[\frac{1}{M_2}\sum_{m_2=1}^{M_2} f(y_i|x_{1i},x_{2im_2},x_{3i},\theta)\right]. \end{aligned} \tag{5.30}$$

d) A observação i tem valores faltantes em X_1 e X_2.
Como vimos nos cálculos da Seção 4.1, como temos X_1 e X_2 com valores faltantes, precisamos integrar $f(y_i,x_{1i},x_{2i}|x_{3i},\theta)$ em relação a x_{1i} e x_{2i}, sendo $f(y_i,x_{1i},x_{2i}|x_{3i},\theta)$ dada por:

$$f(y_i,x_{1i},x_{2i}|x_{3i},\theta) = f(y_i|x_{1i},x_{2i},x_{3i},\theta)f(x_{1i}|x_{2i},x_{3i},\theta)f(x_{2i}|x_{3i},\theta). \tag{5.31}$$

Logo,

$$\begin{aligned} f(y_i|x_{3i},\theta) &= \int\int f(y_i,x_{1i},x_{2i}|x_{3i},\theta)dx_{1i}dx_{2i} \\ &= \int\int f(y_i|x_{1i},x_{2i},x_{3i},\theta)f(x_{1i}|x_{2i},x_{3i},\theta)f(x_{2i}|x_{3i},\theta)dx_{1i}dx_{2i} \\ &= \int E_{X_1|X_2,X_3}[f(y_i|X_{1i},x_{2i},x_{3i},\theta)]f(x_{2i}|x_{3i},\theta)dx_{2i} \\ &= E_{X_2|X_3}\left[E_{X_1|X_2,X_3}[f(y_i|X_{1i},X_{2i},x_{3i},\theta)]\right]. \end{aligned} \tag{5.32}$$

Seja n_4 ($n_4 < n$) o número de observações que possuem valores faltantes em X_1 e X_2. Então, geramos M_1 valores aleatórios para esta observação em relação a X_1, $x_{1i1},...,x_{1iM_1}$, da distribuição de $X_1|X_2,X_3$ e M_3 valores aleatórios para esta observação em relação a X_2, $x_{2i1},...,x_{2iM_3}$, da distribuição de $X_2|X_3$, para $i = 1,2,...,n_4$, e a aproximação Monte Carlo de (5.32), para esta densidade, se torna:

$$
\begin{aligned}
f(y_i|x_{3i},\theta) &= E_{X_2|X_3}\left[E_{X_1|X_2,X_3}[f(y_i|X_{1i},X_{2i},x_{3i},\theta)]\right] \\
&= E_{X_2|X_3}\left[\frac{1}{M_1}\sum_{m_1=1}^{M_1} f(y_i|x_{1im_1},X_{2i},x_{3i},\theta)\right] \\
&= \frac{1}{M_3}\sum_{m_3=1}^{M_3}\left[\frac{1}{M_1}\sum_{m_1=1}^{M_1} f(y_i|x_{1im_1},x_{2im_3},x_{3i},\theta)\right] \\
&= \frac{1}{M_3}\frac{1}{M_1}\sum_{m_3=1}^{M_3}\sum_{m_1=1}^{M_1} f(y_i|x_{1im_1},x_{2im_3},x_{3i},\theta). \qquad (5.33)
\end{aligned}
$$

Considere $\delta_1(i)$ a função indicadora se a observação i é observada em relação a X_1 e $\delta_2(i)$ a função indicadora se a observação i é observada em relação a X_2, conforme definição na Seção 4.1. A função de verossimilhança completa, considerando todas as configurações possíveis para a observação i, é escrita como:

$$
\begin{aligned}
L(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) = \prod_{i=1}^{n} & \Big[[f(y_i|x_{1i},x_{2i},x_{3i},\theta)f(x_{1i}|x_{2i},x_{3i},\theta)f(x_{2i}|x_{3i},\theta)]^{\delta_1(i)\delta_2(i)} \\
& \times [f(y_i|x_{2i},x_{3i},\theta)]^{(1-\delta_1(i))\delta_2(i)} \\
& \times [f(y_i|x_{1i},x_{3i},\theta)]^{(1-\delta_2(i))\delta_1(i)} \\
& \times [f(y_i|x_{3i},\theta)]^{(1-\delta_1(i))(1-\delta_2(i))}\Big], \qquad (5.34)
\end{aligned}
$$

em que θ é o vetor de parâmetros a ser estimado.

Nosso objetivo é encontrar as estimativas que tornam máximo o valor da função de verossimilhança completa (5.34). Dessa forma, o próximo passo para encontrarmos as estimativas de máxima-verossimilhança de θ é calcularmos a função de log-verossimilhança completa $l(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3)$ correspondente a (5.34):

$$
\begin{aligned}
l(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) &= \log L(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) \\
&= \sum_{i=1}^{n} [\delta_1(i)\delta_2(i)(\log[f(y_i|x_{1i},x_{2i},x_{3i},\theta)f(x_{1i}|x_{2i},x_{3i},\theta)f(x_{2i}|x_{3i},\theta)]) \\
&+ ((1-\delta_1(i))\delta_2(i)\log[f(y_i|x_{2i},x_{3i},\theta)]) \\
&+ ((1-\delta_2(i))\delta_1(i)\log[f(y_i|x_{1i},x_{3i},\theta)]) \\
&+ ((1-\delta_1(i))(1-\delta_2(i))\log[f(y_i|x_{3i},\theta)])]. \qquad (5.35)
\end{aligned}
$$

Reescrevendo a função de log-verossimilhança completa dada em (5.35) em relações às aproximações Monte Carlo para cada densidade, temos:

$$
\begin{aligned}
l(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) \approx \sum_{i=1}^{n} & [(\delta_1(i)\delta_2(i)\log[f(y_i|x_{1i},x_{2i},x_{3i},\theta)f(x_{1i}|x_{2i},x_{3i},\theta)f(x_{2i}|x_{3i},\theta)]) \\
+ & \left((1-\delta_1(i))\delta_2(i)\log\left[\tfrac{1}{M_1}\sum_{m_1=1}^{M_1} f(y_i|x_{1im_1},x_{2i},x_{3i},\theta)\right]\right) \\
+ & \left((1-\delta_2(i))\delta_1(i)\log\left[\tfrac{1}{M_2}\sum_{m_2=1}^{M_2} f(y_i|x_{1i},x_{2im_2},x_{3i},\theta)\right]\right) \\
+ & \left((1-\delta_1(i))(1-\delta_2(i))\log\left[\tfrac{1}{M_3}\tfrac{1}{M_1}\sum_{m_3=1}^{M_3}\sum_{m_1=1}^{M_1} f(y_i|x_{1im_1},x_{2im_3},x_{3i},\theta)\right]\right)\Bigg],
\end{aligned}
$$

sendo $x_{1i1},\ldots,x_{1iM_1}$ os M_1 valores aleatórios gerados da distribuição de $X_1|X_2,X_3$, $x_{2i1},\ldots,x_{2iM_2}$ os M_2 valores aleatórios gerados da distribuição de $X_2|X_1,X_3$ e $x_{2i1},\ldots,x_{2iM_3}$ os M_3 valores aleatórios gerados da distribuição de $X_2|X_3$. Observe que a distribuição de $X_2|X_1,X_3$ não aparece na função de log-verossimilhança de casos completos e, sendo assim, não conseguimos estimar os parâmetros associados a ela. Dessa maneira, substituímos essa distribuição pela distribuição de $X_2|X_3$ para simular valores de X_2 quando apenas o valor de X_2 é faltante.

5.2.2 Método utilizando média de log-verossimilhanças

Considerando a abordagem utilizando a média de log-verossimilhanças, primeiramente vamos construir as funções de log-verossimilhança referentes aos quatro cenários possíveis para uma observação. São eles:

a) A observação i tem todas as suas variáveis observadas.
Seja n_1 $(n_1 < n)$ o número de observações que não possuem valores faltantes. Neste caso, consideraremos a distribuição de $Y,X_1,X_2|X_3$ para estimação dos parâmetros quando temos todos os valores observados, ou seja:

$$
L(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) = \prod_{i=1}^{n_1} f(y_i|x_{1i},x_{2i},x_{3i},\theta)f(x_{1i}|x_{2i},x_{3i},\theta)f(x_{2i}|x_{3i},\theta). \tag{5.36}
$$

Consequentemente, $l(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) = \log L(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3)$ é dada por:

$$
\begin{aligned}
l(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) &= \sum_{i=1}^{n_1} \log[f(y_i|x_{1i},x_{2i},x_{3i},\theta)f(x_{1i}|x_{2i},x_{3i},\theta)f(x_{2i}|x_{3i},\theta)] \qquad (5.37)\\
&= \sum_{i=1}^{n_1} [\log(f(y_i|x_{1i},x_{2i},x_{3i},\theta))+\log(f(x_{1i}|x_{2i},x_{3i},\theta))+\log(f(x_{2i}|x_{3i},\theta))];
\end{aligned}
$$

b) A observação i tem valor faltante em X_1.
Seja n_2 $(n_2 < n)$ o número de observações que possuem valor faltante apenas em X_1.

Neste caso, estamos interessados no cálculo da média das log-verossimilhanças completas referentes à distribuição $f(y_i|x_{2i},x_{3i},\theta)$, para $i=1,...,n_2$, que é dada por:

$$\begin{aligned}\bar{l}(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) &\approx \frac{1}{M_1}\sum_{m_1=1}^{M_1}\left[\sum_{i=1}^{n_2}\log[f(y_i,x_{1im_1}|x_{2i},x_{3i},\theta)]\right] \\ &= \frac{1}{M_1}\sum_{m_1=1}^{M_1}\left[\sum_{i=1}^{n_2}\log[f(y_i|x_{1im_1},x_{2i},x_{3i},\theta)f(x_{1im_1}|x_{2i},x_{3i},\theta)]\right] \\ &= \frac{1}{M_1}\sum_{m_1=1}^{M_1}\left[\sum_{i=1}^{n_2}[\log(f(y_i|x_{1im_1},x_{2i},x_{3i},\theta))+\log(f(x_{1im_1}|x_{2i},x_{3i},\theta))]\right],\end{aligned} \tag{5.38}$$

sendo $x_{1i1},...,x_{1iM_1}$ os M_1 valores aleatórios para a observação i ($i=1,...,n_2$), gerados da distribuição de $X_1|X_2,X_3$;

c) A observação i tem valor faltante em X_2.
Seja n_3 $(n_3<n)$ o número de observações que possuem valor faltante apenas em X_2. Neste caso, estamos interessados no cálculo da média das log-verossimilhanças completas referentes à distribuição $f(y_i|x_{1i},x_{3i},\theta)$, para $i=1,...,n_3$, que é dada por:

$$\begin{aligned}\bar{l}(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) &\approx \frac{1}{M_2}\sum_{m_2=1}^{M_2}\left[\sum_{i=1}^{n_3}\log[f(y_i,x_{2im_2}|x_{1i},x_{3i},\theta)]\right] \\ &= \frac{1}{M_2}\sum_{m_2=1}^{M_2}\left[\sum_{i=1}^{n_3}\log[f(y_i|x_{1i},x_{2im_2},x_{3i},\theta)f(x_{2im_2}|x_{1i},x_{3i},\theta)]\right] \\ &= \frac{1}{M_2}\sum_{m_2=1}^{M_2}\left[\sum_{i=1}^{n_3}[\log(f(y_i|x_{1i},x_{2im_2},x_{3i},\theta))+\log(f(x_{2im_2}|x_{1i},x_{3i},\theta))]\right],\end{aligned} \tag{5.39}$$

sendo $x_{2i1},...,x_{2iM_2}$ os M_2 valores aleatórios para a observação i ($i=1,...,n_3$), gerados da distribuição de $X_2|X_1,X_3$;

d) A observação i tem valores faltantes em X_1 e X_2.
Seja n_4 $(n_4<n)$ o número de observações que possuem valor faltante em X_1 e X_2. Neste caso, estamos interessados no cálculo da média das log-verossimilhanças completas referentes à distribuição $f(y_i|x_{3i},\theta)$, para $i=1,...,n_4$, que é dada por:

$$
\begin{aligned}
\bar{l}(\boldsymbol{\theta}|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) &\approx \frac{1}{M_3}\sum_{m_3=1}^{M_3}\frac{1}{M_1}\sum_{m_1=1}^{M_1}\left[\sum_{i=1}^{n_4}\log[f(y_i,x_{1im_1},x_{2im_3}|x_{3i},\boldsymbol{\theta})]\right] \\
&= \frac{1}{M_3}\sum_{m_3=1}^{M_3}\frac{1}{M_1}\sum_{m_1=1}^{M_1}\left[\sum_{i=1}^{n_4}\log\big[f(y_i|x_{1im_1},x_{2im_3},x_{3i},\boldsymbol{\theta})f(x_{1im_1}|x_{2im_3},x_{3i},\boldsymbol{\theta})\right. \\
&\quad\times\quad \left. f(x_{2im_3}|x_{3i},\boldsymbol{\theta})\big]\right] \\
&= \frac{1}{M_3}\sum_{m_3=1}^{M_3}\frac{1}{M_1}\sum_{m_1=1}^{M_1}\left[\sum_{i=1}^{n_4}\big[\log(f(y_i|x_{1im_1},x_{2im_3},x_{3i},\boldsymbol{\theta}))\right. \\
&\quad+\quad \left.\log(f(x_{1im_1}|x_{2im_3},x_{3i},\boldsymbol{\theta}))+\log(f(x_{2im_3}|x_{3i},\boldsymbol{\theta}))\big]\right] \\
&= \frac{1}{M_3}\frac{1}{M_1}\sum_{m_3=1}^{M_3}\sum_{m_1=1}^{M_1}\left[\sum_{i=1}^{n_4}\big[\log(f(y_i|x_{1im_1},x_{2im_3},x_{3i},\boldsymbol{\theta}))\right. \\
&\quad+\quad \left.\log(f(x_{1im_1}|x_{2im_3},x_{3i},\boldsymbol{\theta}))+\log(f(x_{2im_3}|x_{3i},\boldsymbol{\theta}))\big]\right],
\end{aligned}
\tag{5.40}
$$

sendo $x_{1i1},...,x_{1iM_1}$ os M_1 valores aleatórios para a observação i ($i = 1,...,n_4$), gerados da distribuição de $X_1|X_2,X_3$ e $x_{2i1},...,x_{2iM_3}$ os M_3 valores aleatórios para a observação gerados da distribuição de $X_2|X_3$.

Considere $\delta_1(i)$ a função indicadora se a observação i é observada em relação a X_1 e $\delta_2(i)$ a função indicadora se a observação i é observada em relação a X_2. A média das log-verossimilhanças completas $\bar{l}(\boldsymbol{\theta}|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3)$, estendida para os quatro cenários possíveis para a observação i, é definida como:

$$
\begin{aligned}
\bar{l}(\boldsymbol{\theta}|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) \approx \frac{1}{M_3}\sum_{m_3=1}^{M_3}\ & \frac{1}{M_2}\sum_{m_2=1}^{M_2}\frac{1}{M_1}\sum_{m_1=1}^{M_1}\left[\sum_{i=1}^{n}[\delta_1(i)\delta_2(i)([\log(f(y_i|x_{1i},x_{2i},x_{3i},\boldsymbol{\theta}))\right. \\
&+ \left.\log(f(x_{1i}|x_{2i},x_{3i},\boldsymbol{\theta}))+\log(f(x_{2i}|x_{3i},\boldsymbol{\theta}))])\right. \\
&+ (1-\delta_1(i))\delta_2(i)\big(\big[\log(f(y_i|x_{1im_1},x_{2i},x_{3i},\boldsymbol{\theta}))+\log(f(x_{1im_1}|x_{2i},x_{3i},\boldsymbol{\theta}))\big]\big) \\
&+ (1-\delta_2(i))\delta_1(i)\big(\big[\log(f(y_i|x_{1i},x_{2im_2},x_{3i},\boldsymbol{\theta}))+\log(f(x_{2im_2}|x_{1i},x_{3i},\boldsymbol{\theta}))\big]\big) \\
&+ (1-\delta_1(i))(1-\delta_2(i))\big(\big[\log(f(y_i|x_{1im_1},x_{2im_3},x_{3i},\boldsymbol{\theta}))\big. \\
&+ \left.\big.\log(f(x_{1im_1}|x_{2im_3},x_{3i},\boldsymbol{\theta}))+\log(f(x_{2im_3}|x_{3i},\boldsymbol{\theta}))\big]\big)\big]\big],
\end{aligned}
$$

sendo $x_{1i1},...,x_{1iM_1}$ os M_1 valores aleatórios geradas da distribuição de $X_1|X_2,X_3$, $x_{2i1},...,x_{2iM_2}$ os M_2 valores aleatórios gerados da distribuição de $X_2|X_1,X_3$ e $x_{2i1},...,x_{2iM_3}$ os M_3 valores aleatórios gerados da distribuição de $X_2|X_3$. Observe que a distribuição de $X_2|X_1,X_3$ não aparece na função de log-verossimilhança de casos completos e, sendo assim, não conseguimos estimar os parâmetros associados a ela. Dessa maneira, substituímos essa distribuição pela distribuição de $X_2|X_3$ para simular valores de X_2 quando apenas o valor de X_2 é faltante.

5.2.3 Método utilizando o algoritmo EM

Consideramos agora o método de estimação baseado no algoritmo EM. Os quatro cenários possíveis para cada observação i são:

a) A observação i tem todas as suas variáveis observadas.
Seja n_1 $(n_1 < n)$ o número de observações que não possuem valores faltantes. Neste caso, consideraremos a distribuição de $Y, X_1, X_2|X_3$ para estimação dos parâmetros quando temos todos os valores observados, ou seja:

$$L(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) = \prod_{i=1}^{n_1} f(y_i|x_{1i},x_{2i},x_{3i},\theta) f(x_{1i}|x_{2i},x_{3i},\theta) f(x_{2i}|x_{3i},\theta). \tag{5.41}$$

Consequentemente, $l(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) = \log L(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3)$ é dada por:

$$\begin{aligned} l(\theta|\mathbf{y},\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) &= \sum_{i=1}^{n_1} \log[f(y_i|x_{1i},x_{2i},x_{3i},\theta) f(x_{1i}|x_{2i},x_{3i},\theta) f(x_{2i}|x_{3i},\theta)] \\ &= \sum_{i=1}^{n_1} [\log(f(y_i|x_{1i},x_{2i},x_{3i},\theta)) + \log(f(x_{1i}|x_{2i},x_{3i},\theta)) + \log(f(x_{2i}|x_{3i},\theta))]; \end{aligned} \tag{5.42}$$

b) A observação i tem valor faltante em X_1.
Seja n_2 $(n_2 < n)$ o número de observações que possuem valor faltante apenas em X_1. De maneira análoga à Seção 5.1.3 e considerando apenas as observações com X_1 faltante,

$$\begin{aligned} Q(\theta|\theta_r) &= E_{\mathbf{X}_1|\mathbf{Y},\mathbf{X}_2,\mathbf{X}_3,\theta_r}[l(\theta|Y,X_1,X_2,X_3)|Y=y,X_2=x_2,X_3=x_3,\theta_r] \\ &\approx \frac{1}{M_1} \sum_{m_1=1}^{M_1} \left[\sum_{i=1}^{n_2} k_{rim_1} \left[\log(f(y_i|x_{1im_1}^{(r)},x_{2i},x_{3i},\theta_r)) + \log(f(x_{1im_1}^{(r)}|x_{2i},x_{3i},\theta_r)) \right] \right], \end{aligned} \tag{5.43}$$

em que $k_{rim_1} = \dfrac{f(x_{1im_1}^{(r)}|y_i,x_{2i},x_{3i},\theta_r)}{f(x_{1im_1}^{(r)}|x_{2i},x_{3i},\theta_r)}$ são os pesos da amostragem por importância e $x_{1i1}^{(r)},\ldots,x_{1iM_1}^{(r)}$, para $i = 1,2,\ldots,n_2$, são gerados da distribuição conhecida de $X_1|X_2,X_3$.

O peso k_{rim_1} pode ser aproximado por:

$$k_{rim_1} \approx \frac{f(y_i|x_{1im_1}^{(r)},x_{2i},x_{3i},\theta_r)}{\frac{1}{M_1}\sum_{m_1=1}^{M_1} f(y_i|x_{1im_1}^{(r)},x_{2i},x_{3i},\theta_r)}; \tag{5.44}$$

c) A observação i tem valor faltante em X_2.
Seja n_3 $(n_3 < n)$ o número de observações que possuem valor faltante apenas em X_2

e seja M_2 o número de valores aleatórios para X_{2i}, $x_{2i1}^{(r)},...,x_{2iM_2}^{(r)}$ para $i = 1,2,...,n_3$, da distribuição de $X_2|Y,X_1,X_3$.

Então, a aproximação de Monte Carlo para Q, quando as observações são faltantes apenas em X_2, de forma análoga a que vimos na Seção 5.1.3, é:

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}_r) &= E_{\mathbf{X}_1|\mathbf{Y},\mathbf{X}_2,\mathbf{X}_3,\boldsymbol{\theta}_r}[l(\boldsymbol{\theta}|Y,X_1,X_2,X_3)|Y=y,X_1=x_1,X_3=x_3,\boldsymbol{\theta}_r] \\ &\approx \frac{1}{M_2}\sum_{m_2=1}^{M_2}\left[\sum_{i=1}^{n_3} k_{rim_2}\left[\log(f(y_i|x_{1i},x_{2im_2}^{(r)},x_{3i},\boldsymbol{\theta}_r))+\log(f(x_{2im_2}^{(r)}|x_{1i},x_{3i},\boldsymbol{\theta}_r))\right]\right], \quad (5.45) \end{aligned}$$

em que $k_{rim_2} = \frac{f(x_{2im_2}^{(r)}|y_i,x_{1i},x_{3i},\boldsymbol{\theta}_r)}{f(x_{2im_2}^{(r)}|x_{1i},x_{3i},\boldsymbol{\theta}_r)}$ são os pesos da amostragem por importância e $x_{2i1}^{(r)},...,x_{2iM_2}^{(r)}$, com $i = 1,2,...,n_3$, são gerados da distribuição conhecida de $X_2|X_1,X_3$.

O peso k_{rm_2i} pode ser aproximado por:

$$k_{rim_2} \approx \frac{f(y_i|x_{1i},x_{2im_2}^{(r)},x_{3i},\boldsymbol{\theta}_r)}{\frac{1}{M_2}\sum_{m_2=1}^{M_2} f(y_i|x_{1i},x_{2im_2}^{(r)},x_{3i},\boldsymbol{\theta}_r)}; \quad (5.46)$$

d) A observação i tem valores faltantes em X_1 e X_2.
Seja n_4 $(n_4 < n)$ o número de observações que possuem valor faltante em X_1 e X_2. Para isso, temos:

$$
\begin{aligned}
q_i(\boldsymbol{\theta}|\boldsymbol{\theta}_r) &= E_{\mathbf{X}_1|\mathbf{Y},\mathbf{X}_2,\mathbf{X}_3,\boldsymbol{\theta}_r}[l(\boldsymbol{\theta}|Y_i,X_{1i},X_{2i},X_{3i})|y_i,x_{3i},\boldsymbol{\theta}_r] \qquad (5.47)\\
&= \int\int l(\boldsymbol{\theta}|x_{1i},x_{2i},x_{3i}) f(x_{1i},x_{2i}|y_i,x_{3i},\boldsymbol{\theta}_r)dx_{1i}dx_{2i}\\
&= \int\int \frac{l(\boldsymbol{\theta}|x_{1i},x_{2i},x_{3i}) f(x_{1i},x_{2i}|y_i,x_{3i},\boldsymbol{\theta}_r) f(x_{1i}|x_{2i},x_{3i},\boldsymbol{\theta}_r) f(x_{2i}|x_{3i},\boldsymbol{\theta}_r)}{f(x_{1i}|x_{2i},x_{3i},\boldsymbol{\theta}_r) f(x_{2i}|x_{3i},\boldsymbol{\theta}_r)} dx_{1i}dx_{2i}\\
&= \int\int \left[\frac{l(\boldsymbol{\theta}|x_{1i},x_{2i},x_{3i}) f(x_{1i},x_{2i}|y_i,x_{3i},\boldsymbol{\theta}_r) f(x_{1i}|x_{2i},x_{3i},\boldsymbol{\theta}_r)}{f(x_{1i}|x_{2i},x_{3i},\boldsymbol{\theta}_r) f(x_{2i}|x_{3i},\boldsymbol{\theta}_r)} dx_{1i}\right] f(x_{2i}|x_{3i},\boldsymbol{\theta}_r)dx_{2i}\\
&= \int E_{X_1|X_2,X_3}\left[\frac{l(\boldsymbol{\theta}|X_{1i},x_{2i},x_{3i}) f(X_{1i},x_{2i}|y_i,x_{3i},\boldsymbol{\theta}_r)}{f(X_{1i}|x_{2i},x_{3i},\boldsymbol{\theta}_r) f(x_{2i}|x_{3i},\boldsymbol{\theta}_r)}\right] f(x_{2i}|x_{3i},\boldsymbol{\theta}_r)dx_{2i}\\
&= E_{X_2|X_3}\left[E_{X_1|X_2,X_3}\left[\frac{l(\boldsymbol{\theta}|X_{1i},X_{2i},x_{3i}) f(X_{1i},X_{2i}|y_i,x_{3i},\boldsymbol{\theta}_r)}{f(X_{1i}|X_{2i},x_{3i},\boldsymbol{\theta}_r) f(X_{2i}|x_{3i},\boldsymbol{\theta}_r)}\right]\right]\\
&\approx E_{X_2|X_3}\left[\frac{1}{M_1}\sum_{m_1=1}^{M_1}\left[\frac{l(\boldsymbol{\theta}|x^{(r)}_{1im_1},X_{2i},x_{3i}) f(x^{(r)}_{1im_1},X_{2i}|y_i,x_{3i},\boldsymbol{\theta}_r)}{f(x_{1im_1}|X_{2i},x_{3i},\boldsymbol{\theta}_r) f(X_{2i}|x_{3i},\boldsymbol{\theta}_r)}\right]\right]\\
&\approx \frac{1}{M_3}\sum_{m_3=1}^{M_3}\frac{1}{M_1}\sum_{m_1=1}^{M_1}\left[\frac{l(\boldsymbol{\theta}|x^{(r)}_{1im_1},x^{(r)}_{2im_3},x_{3i}) f(x^{(r)}_{1im_1},x^{(r)}_{2im_3}|y_i,x_{3i},\boldsymbol{\theta}_r)}{f(x^{(r)}_{1im_1}|x^{(r)}_{2im_3},x_{3i},\boldsymbol{\theta}_r) f(x^{(r)}_{2im_3}|x_{3i},\boldsymbol{\theta}_r)}\right]\\
&= \frac{1}{M_3}\sum_{m_3=1}^{M_3}\frac{1}{M_1}\sum_{m_1=1}^{M_1}\left[k_{rim_1m_3} l(\boldsymbol{\theta}|x^{(r)}_{1im_1},x^{(r)}_{2im_3},x_{3i})\right]\\
&= \frac{1}{M_3}\sum_{m_3=1}^{M_3}\frac{1}{M_1}\sum_{m_1=1}^{M_1}\Big[k_{rim_1m_3}[\log(f(y_i|x^{(r)}_{1im_1},x^{(r)}_{2im_3},x_{3i},\boldsymbol{\theta}))\\
&\quad + \log(f(x^{(r)}_{1im_1}|x^{(r)}_{2im_3},x_{3i},\boldsymbol{\theta}))+\log(f(x^{(r)}_{2im_3}|x_{3i},\boldsymbol{\theta}))]\Big],
\end{aligned}
$$

em que $k_{rim_1m_3} = \dfrac{f(x^{(r)}_{1im_1},x^{(r)}_{2im_3}|y_i,x_{3i},\boldsymbol{\theta}_r)}{f(x^{(r)}_{1im_1}|x^{(r)}_{2im_3},x_{3i},\boldsymbol{\theta}_r) f(x^{(r)}_{2im_3}|x_{3i},\boldsymbol{\theta}_r)}$ são os pesos da amostragem por importância, $x^{(r)}_{1i1},...,x^{(r)}_{1iM_1}$ são gerados da distribuição conhecida de $X_1|X_2,X_3$ e $x^{(r)}_{2i1},...,x^{(r)}_{2iM_3}$, com $i = 1,2,...,n_4$, são gerados da distribuição conhecida de $X_2|X_3$.

A função $k_{rim_1m_3}$ pode ainda ser simplificada:

$$
\begin{aligned}
k_{rim_1m_3} &= \frac{f(x_{1im_1}^{(r)}, x_{2im_3}^{(r)} | y_i, x_{3i}, \theta_r)}{f(x_{1im_1}^{(r)} | x_{2im_3}^{(r)}, x_{3i}, \theta_r) f(x_{2im_3}^{(r)} | x_{3i}, \theta_r)} \\
&= \frac{f(x_{1im_1}^{(r)}, x_{2im_3}^{(r)}, y_i, x_{3i} | \theta_r)}{f(y_i, x_{3i} | \theta_r) f(x_{1im_1}^{(r)} | x_{2im_3}^{(r)}, x_{3i}, \theta_r) f(x_{2im_3}^{(r)} | x_{3i}, \theta_r)} \\
&= \frac{f(y_i | x_{1im_1}^{(r)}, x_{2im_3}^{(r)}, x_{3i}, \theta_r) f(x_{1im_1}^{(r)}, x_{2im_3}^{(r)}, x_{3i} | \theta_r)}{f(y_i | x_{3i}, \theta_r) f(x_{3i} | \theta_r) f(x_{1im_1}^{(r)} | x_{2im_3}^{(r)}, x_{3i}, \theta_r) f(x_{2im_3}^{(r)} | x_{3i}, \theta_r)} \\
&= \frac{f(y_i | x_{1im_1}^{(r)}, x_{2im_3}^{(r)}, x_{3i}, \theta_r) f(x_{1im_1}^{(r)} | x_{2im_3}^{(r)}, x_{3i}, \theta_r) f(x_{2im_3}^{(r)}, x_{3i} | \theta_r)}{f(y_i | x_{3i}, \theta_r) f(x_{3i} | \theta_r) f(x_{1im_1}^{(r)} | x_{2im_3}^{(r)}, x_{3i}, \theta_r) f(x_{2im_3}^{(r)} | x_{3i}, \theta_r)} \\
&= \frac{f(y_i | x_{1im_1}^{(r)}, x_{2im_3}^{(r)}, x_{3i}, \theta_r) f(x_{1im_1}^{(r)} | x_{2im_3}^{(r)}, x_{3i}, \theta_r) f(x_{2im_3}^{(r)} | x_{3i}, \theta_r) f(x_{3i} | \theta_r)}{f(y_i | x_{3i}, \theta_r) f(x_{3i} | \theta_r) f(x_{1im_1}^{(r)} | x_{2im_3}^{(r)}, x_{3i}, \theta_r) f(x_{2im_3}^{(r)} | x_{3i}, \theta_r)} \\
&= \frac{f(y_i | x_{1im_1}^{(r)}, x_{2im_3}^{(r)}, x_{3i}, \theta_r)}{f(y_i | x_{3i}, \theta_r)} \\
&= \frac{f(y_i | x_{1im_1}^{(r)}, x_{2im_3}^{(r)}, x_{3i}, \theta_r)}{\int\int f(y_i, x_{1i}, x_{2i} | x_{3i}, \theta_r) dx_{1i} dx_{2i}} \\
&= \frac{f(y_i | x_{1im_1}^{(r)}, x_{2im_3}^{(r)}, x_{3i}, \theta_r)}{\int\int f(y_i | x_{1i}, x_{2i}, x_{3i}, \theta_r) f(x_{1i} | x_{2i}, x_{3i}, \theta_r) f(x_{2i} | x_{3i}, \theta_r) dx_{1i} dx_{2i}}.
\end{aligned}
$$

O peso $k_{rim_1m_3}$ pode ser aproximado por:

$$
k_{rim_1m_3} \approx \frac{f(y_i | x_{1im_1}^{(r)}, x_{2im_3}^{(r)}, x_{3i}, \theta_r)}{\frac{1}{M_3} \sum_{m_3=1}^{M_3} \frac{1}{M_1} \sum_{m_1=1}^{M_1} f(y_i | x_{1im_1}^{(r)}, x_{2im_3}^{(r)}, x_{3i}, \theta_r)}. \tag{5.48}
$$

Considere $\delta_1(i)$ a função indicadora se a observação i é observada em relação a X_1 e $\delta_2(i)$ a função indicadora se a observação i é observada em relação a X_2. A expressão $Q(\theta|\theta_r)$ estendida para os quatro cenários possíveis para a observação i, é definida como:

$$
\begin{aligned}
Q(\theta|\theta_r) \approx \frac{1}{M_3}\sum_{m_3=1}^{M_3} & \frac{1}{M_2}\sum_{m_2=1}^{M_2}\frac{1}{M_1}\sum_{m_1=1}^{M_1}\left[\sum_{i=1}^{n}[\delta_1(i)\delta_2(i)\,(\log(f(y_i|x_{1i},x_{2i},x_{3i},\theta)) \right. \\
+ & \log(f(x_{1i}|x_{2i},x_{3i},\theta))+\log(f(x_{2i}|x_{3i},\theta))) \\
+ & (1-\delta_1(i))\delta_2(i)\left(k_{rim_1}\left[\log(f(y_i|x^{(r)}_{1im_1},x_{2i},x_{3i},\theta_r))\right.\right. \\
+ & \left.\left.\log(f(x^{(r)}_{1im_1}|x_{2i},x_{3i},\theta_r))\right]\right) \\
+ & \delta_1(i)(1-\delta_2(i))\left(k_{rim_2}\left[\log(f(y_i|x_{1i},x^{(r)}_{2im_2},x_{3i},\theta_r))\right.\right. \\
+ & \left.\left.\log(f(x^{(r)}_{2im_2}|x_{1i},x_{3i},\theta_r))\right]\right) \\
+ & (1-\delta_1(i))(1-\delta_2(i))\left(k_{rim_1m_3}\left[\log(f(y_i|x^{(r)}_{1im_1},x^{(r)}_{2im_3},x_{3i},\theta))\right.\right. \\
+ & \left.\left.\left.\log(f(x^{(r)}_{1im_1}|x^{(r)}_{2im_3},x_{3i},\theta))+\log(f(x^{(r)}_{2im_3}|x_{3i},\theta))\right]\right)\right]\Bigg],
\end{aligned}
$$

sendo $x_{1i1},\ldots,x_{1iM_1}$ os M_1 valores aleatórios gerados da distribuição de $X_1|X_2,X_3$, $x_{2i1},\ldots,x_{2iM_2}$ os M_2 valores aleatórios gerados da distribuição de $X_2|X_1,X_3$ e $x_{2i1},\ldots,x_{2iM_3}$ os M_3 valores aleatórios gerados da distribuição de $X_2|X_3$. Observe que a distribuição de $X_2|X_1,X_3$ não aparece na função de log-verossimilhança de casos completos e, sendo assim, não conseguimos estimar os parâmetros associados a ela. Dessa maneira, substituímos essa distribuição pela distribuição de $X_2|X_3$ para simular valores de X_2 quando apenas o valor de X_2 é faltante.

5.2.4 Análise preditiva dos métodos

Para analisarmos o desempenho preditivo dos métodos propostos, dividimos o banco de dados em 70% para treino (subconjunto através do qual estimamos os parâmetros) e 30% para teste (subconjunto para o qual calculamos os valores preditos $\widehat{y}$ e comparamos com os observados y). Para o cálculo do $\widehat{y}$ para as observações na base de teste, consideramos o valor esperado estimado da distribuição de $Y|X_1,X_2,X_3$. Logo,

i) Se x_{1i} e x_{2i} são observados, temos:

$$\widehat{y_i} = E[Y_i|x_{1i},x_{2i},x_{3i},\widehat{\theta}] \tag{5.49}$$

em que $\widehat{\theta}$ é o vetor das estimativas dos parâmetros;

ii) Se x_{1i} é faltante e x_{2i} é observado, temos:

$$\widehat{y_i} = \frac{1}{M_1}\sum_{m_1=1}^{M_1} E[Y_i|x_{1im_1},x_{2i},x_{3i},\widehat{\theta}] \tag{5.50}$$

em que $\widehat{\theta}$ é o vetor das estimativas dos parâmetros e, neste cenário, são gerados M_1 valores aleatórios para X_{1i} da distribuição de $X_1|X_2,X_3$;

iii) Se x_{1i} é observado e x_{2i} é faltante, temos:

$$\widehat{y_i} = \frac{1}{M_2} \sum_{m_2=1}^{M_2} E[Y_i | x_{1i}, x_{2im_2}, x_{3i}, \widehat{\theta}] \tag{5.51}$$

em que $\widehat{\theta}$ é o vetor das estimativas dos parâmetros e, neste cenário, são gerados M_2 valores aleatórios para X_{2i} da distribuição de $X_2|X_3$;

iv) Se x_{1i} e x_{2i} são faltantes, temos:

$$\widehat{y_i} = \frac{1}{M_3} \frac{1}{M_1} \sum_{m_3=1}^{M_3} \sum_{m_1=1}^{M_1} E[Y_i | x_{1im_1}, x_{2im_3}, x_{3i}, \widehat{\theta}] \tag{5.52}$$

em que $\widehat{\theta}$ é o vetor das estimativas dos parâmetros e, neste cenário, são gerados M_1 valores aleatórios para X_{1i} da distribuição de $X_1|X_2, X_3$ e M_3 valores aleatórios para X_{2i} da distribuição de $X_2|X_3$.

Efetuado o cálculo do $\widehat{y}$ para a base de teste, calculamos a média das diferenças ao quadrado entre $\widehat{y}$ e y.

5.2.5 Estudo de simulação

Nesta seção, tendo em vista que, para uma variável com valores faltantes, os métodos baseados em modelo utilizando integração por Monte Carlo e média de log-verossimilhanças foram os que apresentaram melhores performances e o por integração por Monte Carlo tem menor tempo de processamento, apresentamos como foi feito o estudo de simulação e comparação do desempenho apenas do método proposto utilizando integração por Monte Carlo com o de outros métodos de imputação e deleção de dados. Nesse cenário com duas variáveis com valores faltantes, também comparamos o desempenho dos diferentes métodos em relação ao viés e ao erro quadrático médio (EQM) das estimativas dos parâmetros e em relação ao poder preditivo definido pela média das diferenças quadráticas entre $\widehat{y}$ e y em amostras teste, separadas especificamente para esse fim, o erro quadrático médio da predição.

Como, tanto para o método com resolução analítica quanto para os métodos sem resolução analítica, as simulações para os mecanismos MCAR e MAR obtiveram resultados semelhantes, além de ambos os mecanismos serem ignoráveis, e como com amostras menores ($n = 100$) já conseguimos diagnosticar o bom funcionamento das metodologias propostas, nesta seção testamos apenas alguns cenários de simulação fixando o tamanho da amostra para $n = 100$ e variando: a proporção de valores faltantes tanto em relação a X_1 quanto em relação a X_2 ($p = 0.20$ e $p = 0.60$) e o mecanismo que gera os dados faltantes, podendo ser MCAR para X_1 e X_2 ou MNAR para X_1 e X_2.

Para cada cenário analisado, simulamos 20 réplicas (amostras diferentes) sob as mesmas condições. Com elas, temos amostras de tamanho 20 para conduzir análises de desempenho

através do viés e EQM das estimativas dos parâmetros, assim como do erro de predição. Quanto mais próximas de zero essas métricas, mais precisa é a estimação e a predição. Aqui, adotamos o modelo Weibull como a distribuição de probabilidades das variáveis, mas qualquer outro modelo pode ser considerado e a metodologia adaptada.

Após a geração do conjunto de dados, separamos as 70% primeiras observações para treino e estimação, através da qual fazemos a análise inferencial dos parâmetros e os outros 30% para teste, em que analisamos o método de estimação baseado em modelo na presença de valores faltantes em relação ao poder preditivo, comparando-o com os métodos: modelo completo (sem dados faltantes), método de deleção *listwise*, método de imputação pela média, método de imputação por *Random Forest*, método de imputação por *hot-deck* e método de imputação múltipla.

Realizamos as simulações de acordo com os seguintes passos:

Passo 1: Geramos x_{3i} da distribuição Normal com média 0 e variância 1, x_{2i} da distribuição Weibull com parâmetro de escala $\exp(\mu_0 + \mu_1 x_{3i})$ e parâmetro de forma α_3, x_{1i} da distribuição Weibull com parâmetro de escala $\exp(\gamma_0 + \gamma_1 x_{2i} + \gamma_2 x_{3i})$ e parâmetro de forma α_2 e y_i da distribuição Weibull com parâmetro de escala $\exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i})$ e parâmetro de forma α_1, para $i = 1, ..., n$, sendo n, portanto, o tamanho da amostra final. Os verdadeiros valores destes parâmetros considerados são: $\alpha_1 = 2,\ \alpha_2 = 2,\ \alpha_3 = 2,\ \beta_0 = 0,\ \beta_1 = 0.3,\ \beta_2 = 0.3,\ \beta_3 = -0.3,\ \gamma_0 = 0,\ \gamma_1 = 0.5,\ \gamma_2 = -0.5,\ \mu_0 = -0.3$ e $\mu_1 = -0.3$;

Passo 2: De acordo com o mecanismo de valores faltantes considerado e o valor de p, $0 \leq p \leq 1$, geramos valores faltantes nas variáveis X_1 e X_2 da mesma forma que descrito na Seção 4.3.2. Lembramos que, nesta seção, o mecanismo MAR não será considerado para simulação e ressaltamos que no caso de estimação pelo modelo completo, para comparação de desempenho, esse passo não é realizado;

Passo 3: Para cada observação i que possui valor faltante apenas em relação a X_1, geramos 50 valores x_{1i} da distribuição de $X_1|X_2, X_3$, ou seja, da distribuição Weibull com parâmetro de forma dado, inicialmente, de maneira não informativa, como 1 e parâmetro de escala $\exp(\gamma_0 + \gamma_1 x_{2i} + \gamma_2 x_{3i})$, com $\gamma_0 = 0,\ \gamma_1 = 0$ e $\gamma_2 = 0$ como valores iniciais; para as observações i que possuem valores faltantes apenas em relação a X_2, geramos 50 valores x_{2i} da distribuição de $X_2|X_3$, ou seja, da distribuição Weibull com parâmetro de forma dado, inicialmente, de maneira não informativa, como 1 e parâmetro de escala $\exp(\mu_0 + \mu_1 x_{3i})$, com $\mu_0 = 0$ e $\mu_1 = 0$ como valores iniciais. Observe que, neste cenário, não estamos gerando x_{2i} da distribuição de $X_2|X_1, X_3$, como foi construído teoricamente na Seção 5.2.1. Por fim, para as observações i que possuem valores faltantes em X_1 e X_2, geramos, primeiramente, 50 valores x_{2i} da distribuição de $X_2|X_3$ e, em seguida, para cada um dos valores gerados x_{2i}, geramos um valor de x_{1i} da distribuição de $X_1|X_2, X_3$, totalizando também 50 valores

para cada x_{1i} faltante (para estas distribuições também assumimos valores iniciais não informativos para os parâmetros);

Passo 4: Maximizamos as funções de log-verossimilhança utilizando o *optim* do software estatístico R com parâmetro *fnscale* $= -1$. O método numérico usado para encontrar o máximo das funções de log-verossimilhança é o de Nelder-Mead. Para obtermos melhores estimativas dos parâmetros, fazemos uma primeira maximização usando como valores iniciais valores não informativos para os parâmetros e, em seguida, maximizamos as funções de log-verossimilhança novamente, também utlizando o método de Nelder-Mead e considerando, como valores iniciais, os valores das estimativas obtidos pelo primeiro processo de maximização. Quanto às funções de log-verossimilhança maximizadas, os procedimentos para os métodos de imputação pela média, imputação por *Random Forest*, imputação por *Hot-Deck*, imputação múltipla, deleção de casos e considerando o conjunto de dados completo, são os mesmos descritos na Seção 4.3.2, porém aqui a função de log-verossimilhança a ser maximizada é:

$$\sum_{i=1}^{n} \log[f(y_i|x_{1i}, x_{2i}, x_{3i}, \alpha_1, \beta_0, \beta_1, \beta_2, \beta_3)],$$

sendo f a densidade da distribuição Weibull com parâmetro de forma α_1 e parâmetro de escala $\exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i})$. Para o método proposto por integração numérica, consideramos a função de log-verossimilhança construída na Seção 5.2.1, em que consideramos os quatro cenários possíveis para cada observação do conjunto de treinamento ou estimação;

Passo 5: Para os métodos aqui propostos sem resolução analítica, repetimos os passos 3 e 4 até a convergência, ou seja, até a diferença entre os valores estimados de cada parâmetro entre uma iteração e a iteração seguinte ser menor do que 10^{-3}, podendo o número máximo de iterações para convergência ser igual a 10000;

Passo 6: Após obtermos os valores das estimativas dos parâmetros para cada conjunto de dados dentro de cada método, calculamos a diferença e a diferença quadrática dessas estimativas em relação aos verdadeiros valores dos parâmetros. Como os parâmetros α_2, α_3, γ_0, γ_1, γ_2, μ_0 e μ_1 são estimados diretamente apenas pelas metodologias propostas, os índices de desempenho dos seus estimadores não são mostrados e comparados;

Passo 7: Para analisarmos os métodos em relação ao poder preditivo, calculamos o erro quadrático médio do $\widehat{y}_i$ em relação ao y_i observado para todas as observações do conjunto de teste. O cálculo de $\widehat{y}_i$ se dá da seguinte forma:

a) Para a metodologia proposta e cada observação i do conjunto de dados de teste, calculamos $\widehat{y}_i$ de acordo com o proposto na Seção 5.2.4. Vale ressaltar que, para cada caso em que a observação i é faltante apenas em relação a X_1, geramos 50 valores de

x_{1i} da distribuição de $X_1|X_2,X_3$, ou seja, da distribuição Weibull com parâmetro de forma dado por $\widehat{\alpha_2}$ (α_2 estimado) e parâmetro de escala $\exp(\widehat{\gamma_0}+\widehat{\gamma_1}x_{2i}+\widehat{\gamma_2}x_{3i})$, sendo $\widehat{\gamma_0}$, $\widehat{\gamma_1}$ e $\widehat{\gamma_2}$ as estimativas dos parâmetros γ_0, γ_1 e γ_2, respectivamente; para cada caso em que a observação i é faltante apenas em relação a X_2, geramos 50 valores de x_{2i} da distribuição de $X_2|X_3$, ou seja, da distribuição Weibull com parâmetro de forma dado por $\widehat{\alpha_3}$ (α_3 estimado) e parâmetro de escala $\exp(\widehat{\mu_0}+\widehat{\mu_1}x_{3i})$, sendo $\widehat{\mu_0}$ e $\widehat{\mu_1}$ as estimativas dos parâmetros μ_0 e μ_1, respectivamente; para cada caso em que a observação i é faltante em relação a X_1 e X_2, geramos, primeiramente, 50 valores de x_{2i} da distribuição de $X_2|X_3$ e, em seguida, para cada valor gerado de x_{2i}, geramos um valor de x_{1i} da distribuição de $X_1|X_2,X_3$, totalizando também 50 valores para cada x_{1i} faltante;

b) Para os métodos de imputação de dados por *Random Forest*, imputação por *Hot-Deck* e imputação múltipla, realizamos a imputação dos dados faltantes na base teste usando os mesmos procedimento descritos na Seção 4.3.2 e, com os dados completos, calculamos $\widehat{y}_i$ de acordo com o item i) da Sessão 5.2.4. Para o caso da imputação múltipla, como criamos cinco conjuntos de dados completos, o erro quadrático médio é dado pela média entre os erros quadráticos médios dos cinco conjuntos gerados;

c) Para o método de imputação pela média, a média das observações não faltantes em X_1 do conjunto de treino é o valor a ser imputado nas observações que possuem valores faltantes no conjunto de teste para a variável X_1. Analogamente, a média das observações não faltantes em X_2 do conjunto de treino é o valor a ser imputado nas observações que possuem valores faltantes no conjunto de teste para a variável X_2. Após esta imputações, calculamos $\widehat{y}_i$ de acordo com o item i) da Sessão 5.2.4;

d) Para o método em que consideramos o conjunto de dados de teste completo, sem valores faltantes, calculamos $\widehat{y}_i$ de acordo com o item i) da Sessão 5.2.4;

e) Para o método de deleção de dados, como deletamos as observações que possuem valores faltantes, não conseguimos calcular o $\widehat{y}_i$ para elas pois não existe, de fato, um processo de imputação ou predição de valores faltantes. Logo, não analisamos o desempenho de predição desse método nas observações do conjunto de teste, por não fazer sentido essa comparação.

As Figuras 73 e 75 mostram o desempenho inferencial dos métodos para os dois cenários estudados em relação ao mecanismo MCAR. As Figuras 74 e 76 mostram os erros quadráticos médios de predição observados para as mesmas simulações.

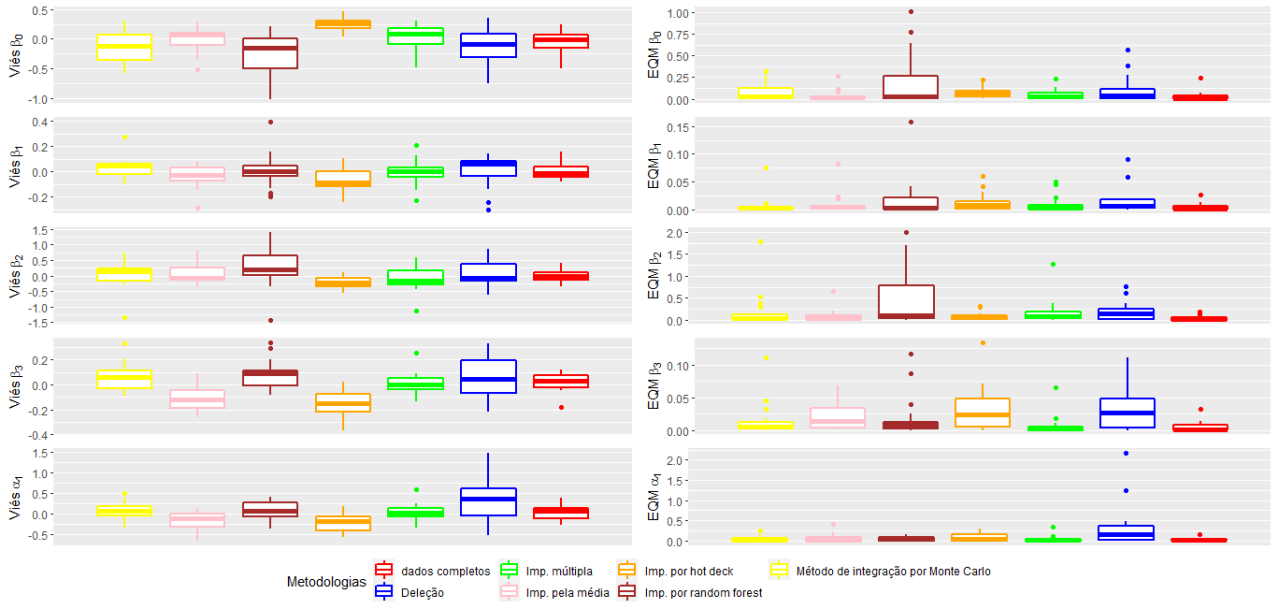

Figura 73 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário com X_1 e X_2 MCAR, $n = 100$, $p = 0.20$ para X_1 e $p = 0.60$ para X_2.

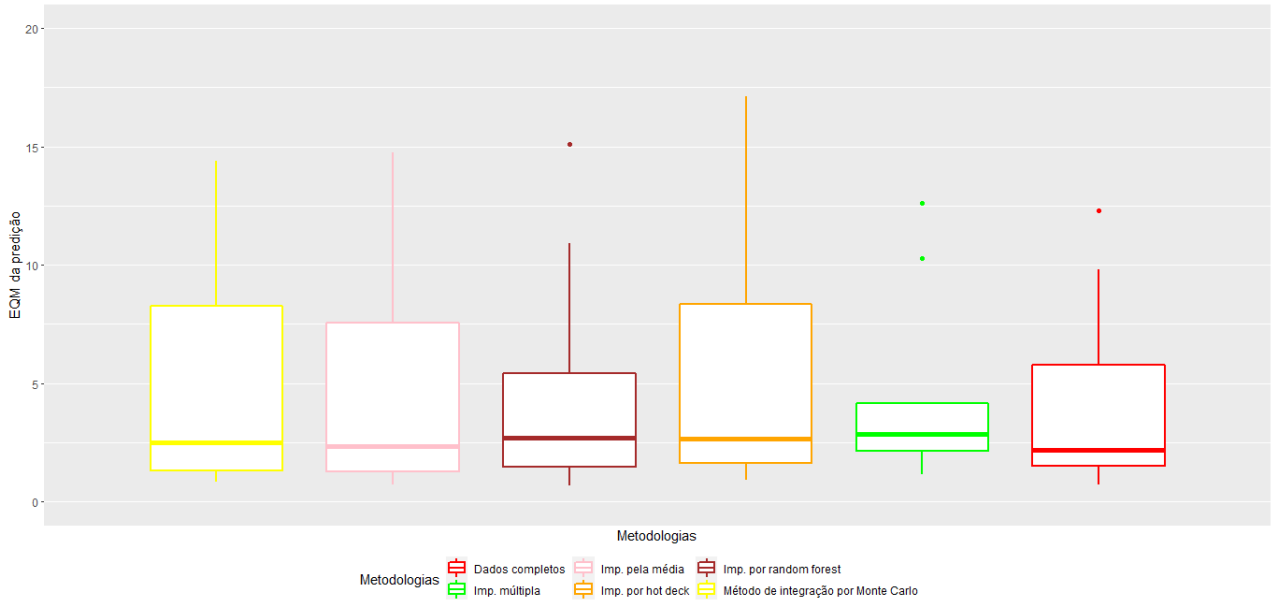

Figura 74 – Erro quadrático médio dos valores preditos para o cenário com X_1 e X_2 MCAR, $n = 100$, $p = 0.20$ para X_1 e $p = 0.60$ para X_2.

Para este cenário MCAR, $n = 100$, $p = 0.20$ para X_1 e $p = 0.60$ para X_2, foram detectados mais *outliers*, além dos já exibidos nas figuras. Eles não constam no gráfico, pois reduziriam consideralmente a escala dos boxplots e dificultaria, com isso, a análise dos desempenhos dos métodos. Seguem os valores dos *outliers* detectados que não constam nos gráficos:

- Método de integração por Monte Carlo: 294.50, 2.219609×10^5;
- Método de imputação pela média: 2.274868×10^5;
- Método de imputação por *random forest:* 117.39, 1.220754×10^5;
- Método de imputação por *hot-deck*: 3.274203×10^5;
- Método de imputação múltipla: 1140.45, 1.221879×10^5, 124.53, 23.03, 209.71, 3.491293×10^5;
- Dados completos: 2.44×10^5.

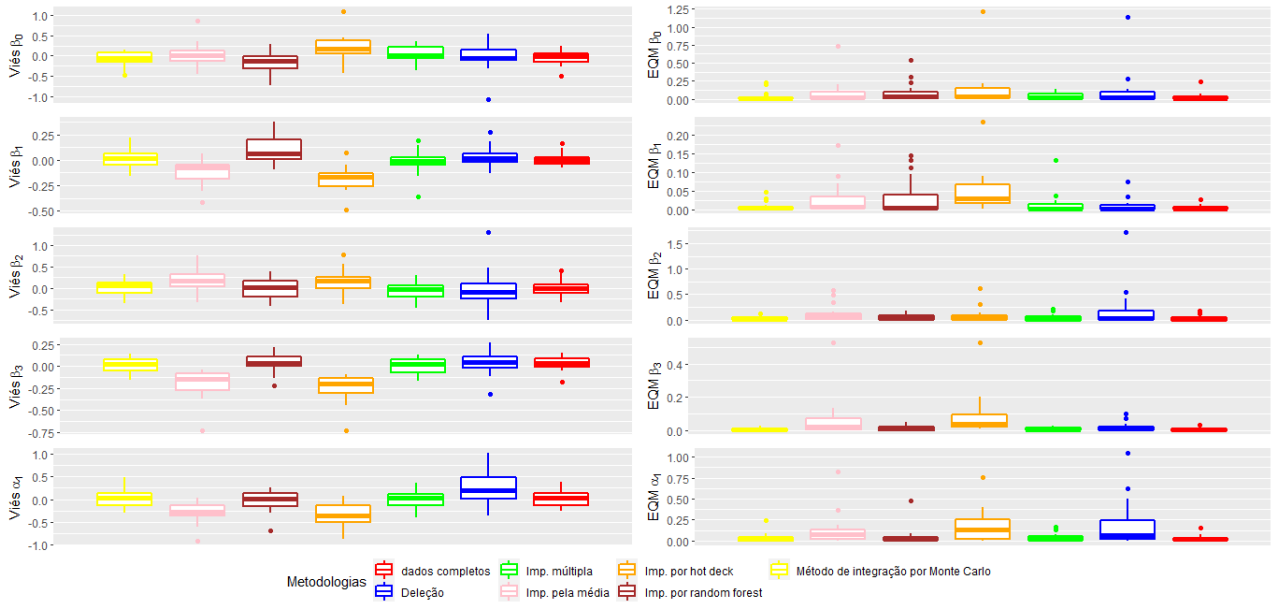

Figura 75 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário com X_1 e X_2 MCAR, $n = 100$, $p = 0.60$ para X_1 e $p = 0.20$ para X_2.

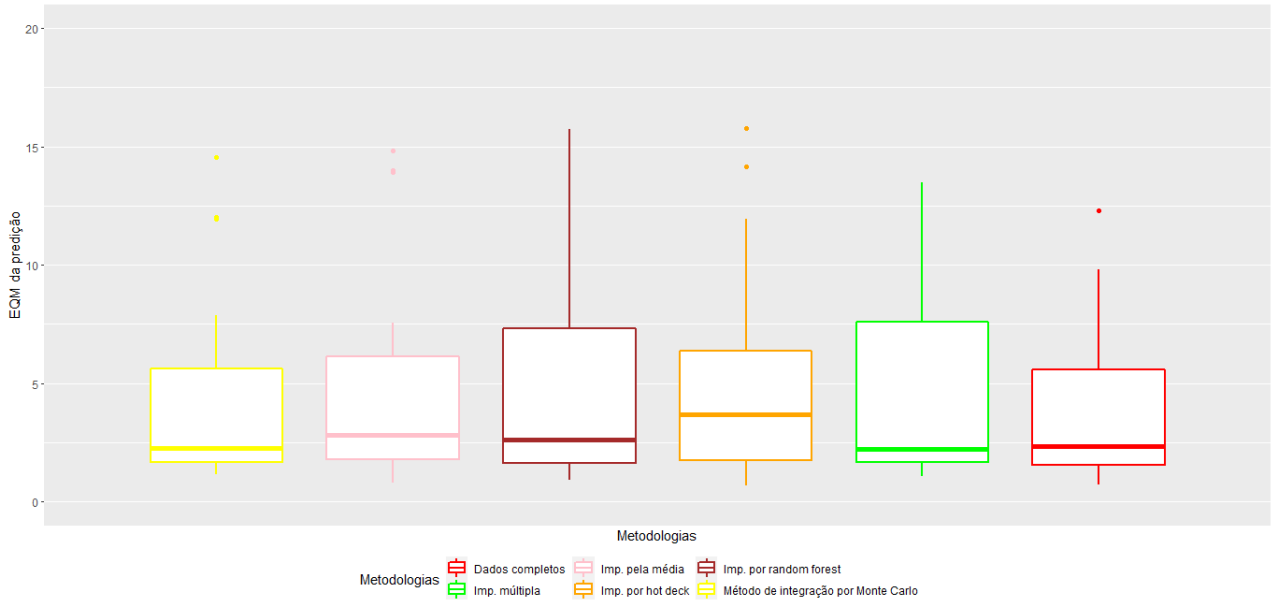

Figura 76 – Erro quadrático médio dos valores preditos para o cenário com X_1 e X_2 MCAR, $n = 100$, $p = 0.60$ para X_1 e $p = 0.20$ para X_2.

Os *outliers* detectados que não constam nos gráficos para o cenário MCAR, $n = 100$, $p = 0.60$ para X_1 e $p = 0.20$ para X_2, são:

- Método de integração por Monte Carlo: 6382.98, 21.94, 2.86×10^7, 3.06×10^5;
- Método de imputação pela média: 23.19, 2.89×10^5;
- Método de imputação por *random forest:* 152.85, 23.53, 9.08×10^4;
- Método de imputação por *hot-deck*: 24.29, 3.90×10^5;
- Método de imputação múltipla: 24.32, 29.29, 42.36, 66.84, 28.85, 38.89, 109.87, 2.16×10^6, 1.25×10^{18};
- Dados completos: 2.44×10^5.

Considerando o mecanismo MCAR de geração de dados faltantes tomando 20% de valores faltantes para X_1 e 60% para X_2, observamos que o método de integração por Monte Carlo, juntamente com o método de imputação pela média, imputação por *Hot-deck*, imputação múltipla e o método considerando os dados completos, obtiveram as melhores performances inferenciais. Por outro lado, no cenário em que consideramos 60% de valores faltantes para X_1 e 20% para X_2, os métodos de integração por Monte Carlo, imputação múltipla e método considerando os dados completos se sobressaem aos demais.

Em relação ao desempenho preditivo nas amostras de teste, o método de imputação por *Random Forest*, imputação múltipla e método considerando os dados completos apresentam resultados melhores do que os do método proposto para o caso em que consideramos 20% de valores faltantes para X_1 e 60% para X_2 em termos de variabilidade. No entanto, a mediana e o primeiro quartil dos EQMs de predição são menores para o método de integração por Monte Carlo, comparável à predição com os dados completos. Já, para o caso em que consideramos 60% de valores faltantes para X_1 e 20% para X_2, o método de integração por Monte Carlo, juntamente com a predição via dados completos, são os que apresentam melhor poder preditivo.

As Figuras 77 e 79 apresentam os resultados inferenciais dos métodos para o mecanismo MNAR de dados faltantes e as Figuras 78 e 80 o desempenho preditivo.

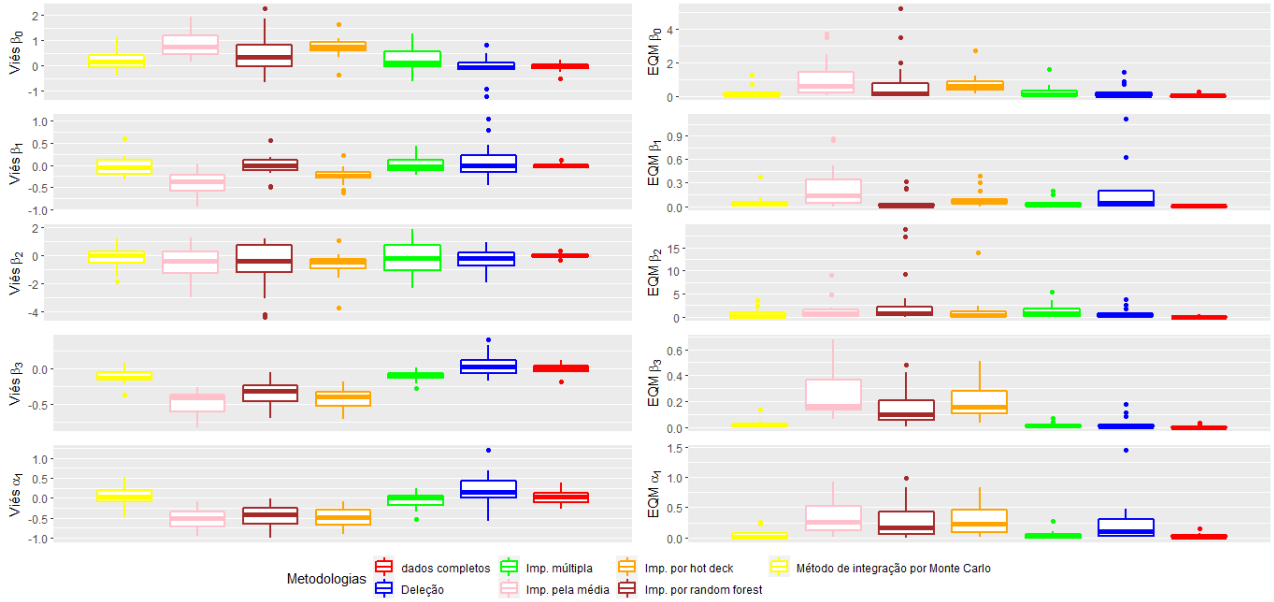

Figura 77 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário com X_1 e X_2 MNAR, $n = 100$, $p = 0.20$ para X_1 e $p = 0.60$ para X_2.

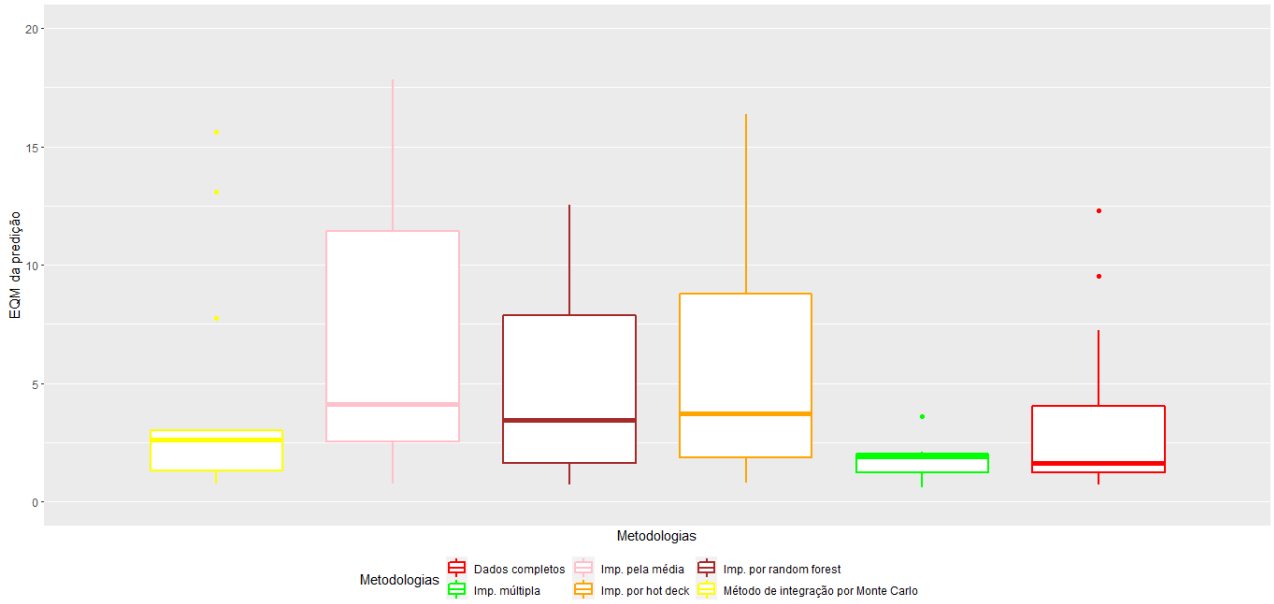

Figura 78 – Erro quadrático médio dos valores preditos para o cenário com X_1 e X_2 MNAR, $n = 100$, $p = 0.20$ para X_1 e $p = 0.60$ para X_2.

Os *outliers* detectados que não constam nos gráficos para o cenário MNAR, $n = 100$, $p = 0.20$ para X_1 e $p = 0.60$ para X_2, são:

- Método de integração por Monte Carlo: 20.56, 25.45;
- Método de imputação por *random forest:* 25.22, 25.73;
- Método de imputação múltipla: 30.04, 67.60, 197.38, 655.71, 6877.48.

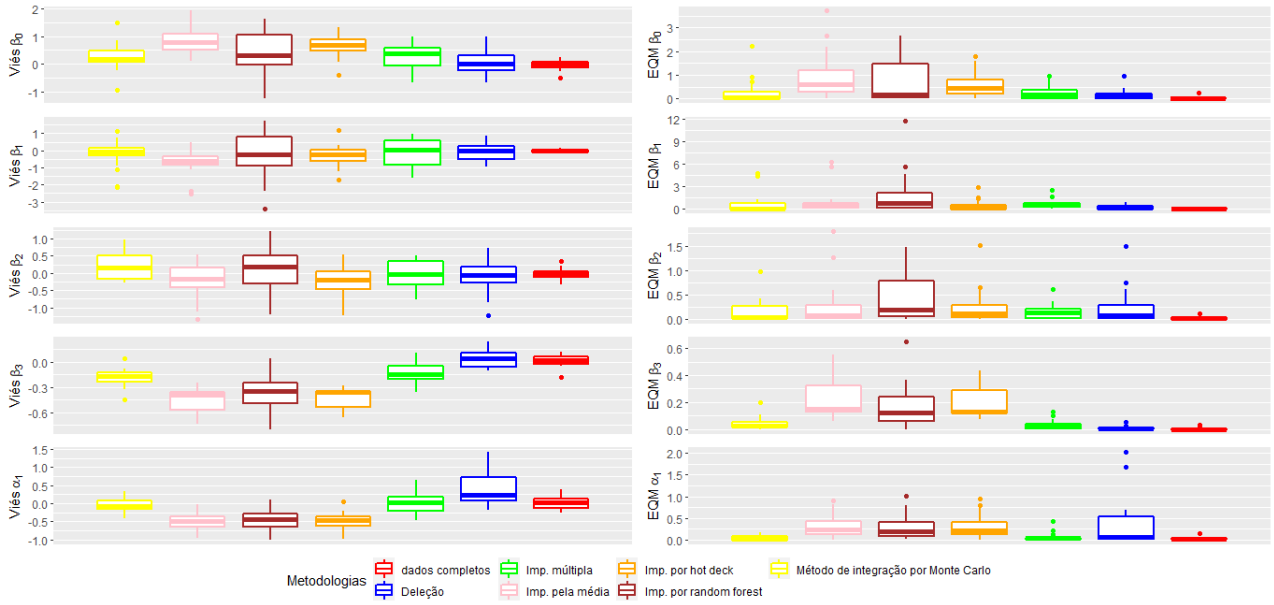

Figura 79 – Viés e erro quadrático médio das estimativas dos parâmetros para o cenário com X_1 e X_2 MNAR, $n = 100$, $p = 0.60$ para X_1 e $p = 0.20$ para X_2.

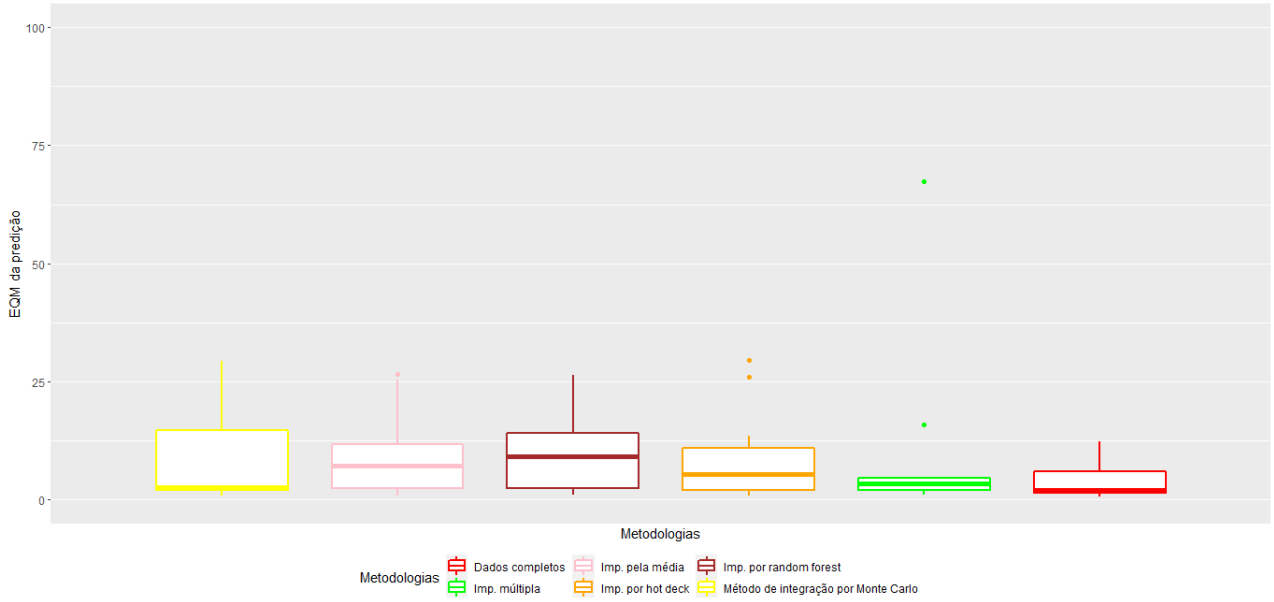

Figura 80 – Erro quadrático médio dos valores preditos para o cenário com X_1 e X_2 MNAR, $n = 100$, $p = 0.60$ para X_1 e $p = 0.20$ para X_2.

Os *outliers* detectados que não constam nos gráficos para o cenário MNAR, $n = 100$, $p = 0.60$ para X_1 e $p = 0.20$ para X_2, apareceram apenas para o método de imputação múltipla. São eles: 6.6710×10^2, 5.3753×10^5, 1.6438×10^2, 2.0694×10^7, 6.012×10^5 e 1.2305×10^{10}.

Considerando o mecanismo MNAR de geração de dados faltantes tomando 20% de valores faltantes para X_1 e 60% para X_2, observamos que o método de integração por Monte Carlo, juntamente com o método de imputação múltipla e o método considerando os dados completos, obtiveram as melhores performances inferenciais. Este comportamento inferencial se mantem para o cenário em que consideramos 60% de valores faltantes para X_1 e 20% para X_2, ressaltando que, em relação ao viés, o método de interação por Monte Carlo apresenta menor variabilidade do que o método de imputação múltipla.

Em relação ao desempenho preditivo nas amostras de teste, os métodos de imputação múltipla, método considerando os dados completos e o método proposto se sobressaem ao demais para o caso em que consideramos 20% de valores faltantes para X_1 e 60% para X_2. No entanto, vale ressaltar que, embora a variabilidade do método de imputação múltipla seja menor que os demais, ele é o que apresenta mais e maiores *outliers*. Para o cenário em que consideramos 60% de valores faltantes para X_1 e 20% para X_2, novamente, os mesmos três métodos se sobressaem, observando que, embora a variabilidade do método de imputação múltipla seja a menor, ele é o único que apresenta *outliers* neste caso, além do fato de a mediana do método baseado em modelo e do método com dados completos serem menores que a do método de imputação múltipla.

CAPÍTULO

6

APLICAÇÃO EM DADOS REAIS

Neste capítulo, apresentamos uma aplicação dos métodos estudados e propostos no conjunto de dados reais sobre condições climáticas e qualidade do ar *Airquality* (CHAMBERS *et al.*, 2018), disponível no *software* estatístico R.

6.1 Dados da qualidade do ar

O conjunto de dados *Airquality* se refere às medições diárias da qualidade do ar em Nova York, de maio a setembro de 1973, armazenadas em um quadro de dados com 153 observações em 6 variáveis. Os dados foram obtidos do Departamento de Conservação do Estado de Nova York (dados de ozônio) e do National Weather Service (dados meteorológicos) e estão publicamente disponíveis no software estatístico R.

As variáveis que compõem esse conjunto de dados são: *Ozone* (medida em ppb), *Solar.R* (medida em lang), *Wind* (medida em mph), *Temp* (medida em graus F), *Month* (mês do ano, variando de 5 a 9, já que as medições foram feitas de maio a setembro) e *Day* (dia do mês, podendo variar de 1 a 31). Vejamos o que representam as variáveis citadas:

a) ***Ozone***: média de ozônio em partes por bilhão de 1300 a 1500 horas na Ilha Roosevelt;

b) ***Solar.R***: radiação solar em *langleys* na faixa de frequência 4000–7700 *angstroms* das 8 horas às 12 horas no Central Park;

c) ***Wind***: velocidade média do vento em milhas por hora às 7 horas e 10 horas no Aeroporto LaGuardia;

d) ***Temp***: temperatura máxima diária em graus Fahrenheit no Aeroporto La Guardia.

As variáveis *Ozone* e *Solar.R* possuem valores faltantes, sendo a proporção de *missings* igual a 0.24 nos valores da variável *Ozone* (37 observações faltantes) e 0.05 nos valores da variável *Solar.R* (7 observações faltantes).

Vale destacar que, nesse conjunto de dados reais, não sabemos se o mecanismo de dados faltantes é MCAR, MAR ou MNAR, e isso não precisa ser necessariamente checado desde que as metodologias de estimação propostas nesse trabalho devem performar bem em qualquer cenário.

A Figura 81 apresenta os box-plots referentes às variáveis quantitativas da base de dados. De acordo com a Figura 81, obtemos informações sobre as variáveis, como mediana, primeiro quartil, terceiro quartil, valores máximo e mínimo, variabilidade, além da presença ou não de *outliers*. A variável *Temp*, por exemplo, apresenta como mediana o valor de 79 graus Fahrenheit, podendo assumir um máximo de 97 graus e um mínimo de 56 graus.

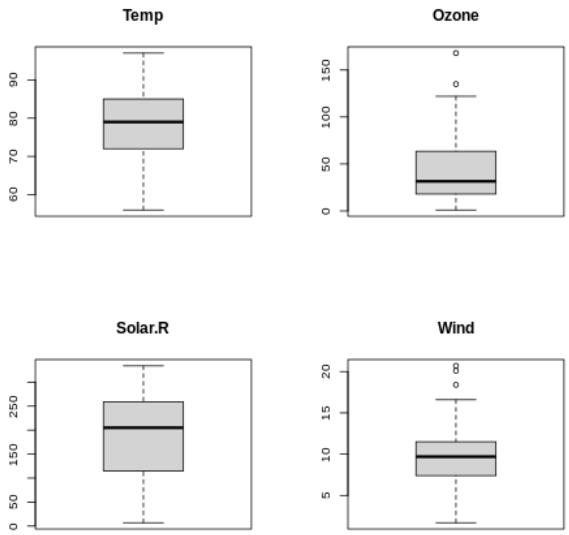

Figura 81 – Box-plots das variáveis quantitativas do dataset *Airquality*.

Analisando agora, por meio de gráficos de dispersão, como varia a temperatura em função das variáveis quantitativas *Ozone*, *Solar.R* e *Wind*, temos a Figura 82. Também podemos construir os gráficos de dispersão de forma a obtermos informações sobre suas distribuições em relação aos meses que foram feitas as medições (de maio a semtembro, variando então de 5 a 9), como mostra a Figura 83. Por elas, obervamos que parece existir uma relação linear entre as variáveis quantitativas e a temperatura.

Apesar dos dados de temperatura, da maneira como foram coletados, se tratarem de uma série temporal, aqui vamos considerar que as observações de temperaturas são independentes condicionadas nos valores das outras variáveis (a serem consideradas como variáveis explicativas).

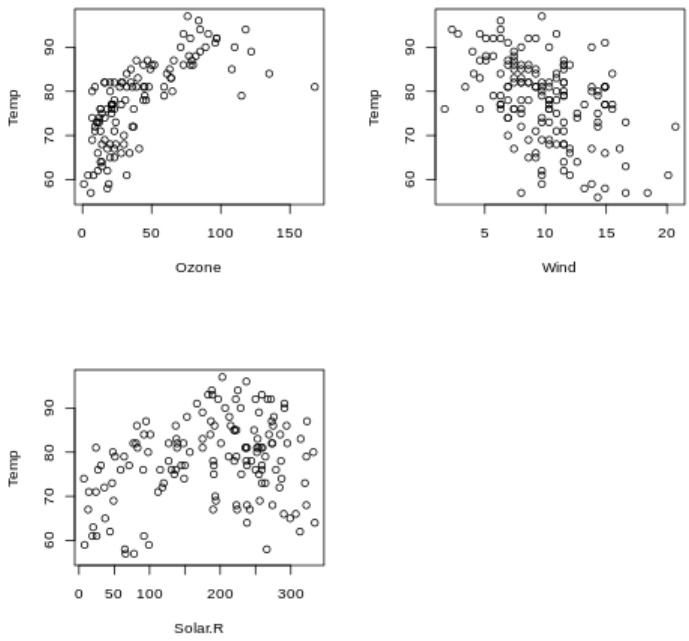

Figura 82 – Gráficos de dispersão da variável *Temp* em função das variáveis quantitativas do dataset *Airquality*.

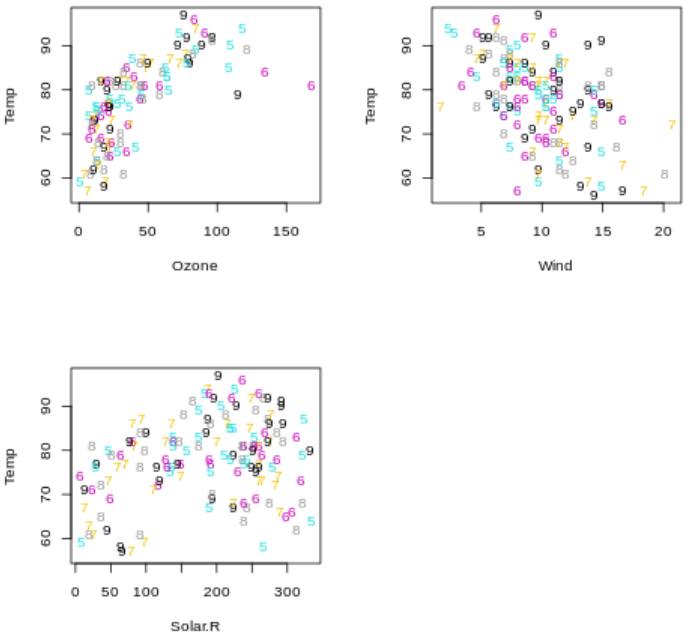

Figura 83 – Gráficos de dispersão da variável *Temp* em função das variáveis quantitativas do dataset *Airquality* discriminadas pela variável *Month* (5-9).

Uma forma gráfica muito utilizada para visualizarmos como as variáveis se relacionam, é o *heatmap* da matriz de correlação. Um *heatmap*, ou mapa de calor, é uma forma de visualização de dados que associa cores à intensidade da correlação entre pares de variáveis analisadas. Cores mais escuras representam correlações mais fortes, enquanto que o azul representa correlação negativa e, o vermelho, correlação positiva.

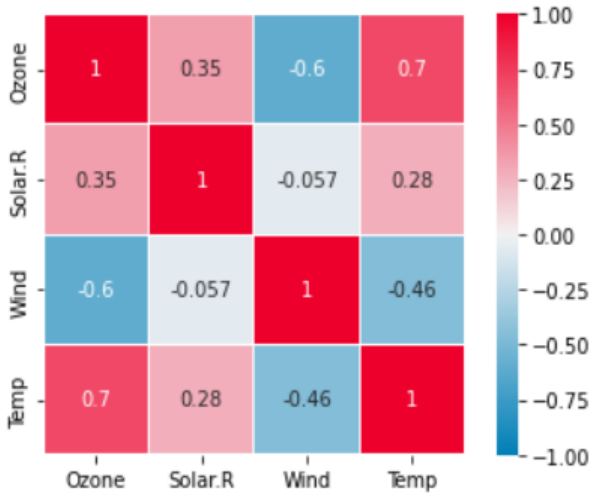

Figura 84 – *Heatmap* das variáveis do dataset *Airquality*.

Pela Figura 84 podemos analisar que, por exemplo, a variável *Temp* tem alta correlação positiva com *Ozone*, ou seja, momentos de alta temperatura tendem a apresentar alta média de ozônio em ppb, e correlação moderada negativa com *Wind*, ou seja, momentos de maior velocidade média de vento apresentam, em geral, temperaturas mais baixas.

6.2 Análise via modelo Gaussiano

Para realizarmos a aplicação da metodologia baseada em modelo com resolução analítica para o dataset *Airquality* e compararmos os resultados com os demais métodos de imputação pela média, imputação por *Random Forest*, imputação por *Hot-deck* e imputação múltipla, realizamos os seguintes procedimentos:

Passo 1: Consideramos *Temp* a variável dependente e *Solar.R, Ozone, Wind* e *Month* as variáveis explicativas. Repare que desconsideramos a variável *Day* do dataset *Airquality* por acreditarmos que o dia do mês não seja importante para predizer a temperatura;

Passo 2: Transformamos a variável categórica *Month*, que assume valores entre 5 e 9 de acordo com o mês a que se refere, em variáveis *dummies*, criando 4 colunas de variáveis dicotômicas, com valores 0 e 1 para os meses 5, 6, 7 e 8 e mantendo o mês 9 como categoria de referência;

Passo 3: Padronizamos os valores das variáveis quantitativas *Temp, Solar.R, Ozone* e *Wind* de modo que todas ficassem centradas no valor zero;

Passo 4: Como os dados são ordenados de acordo com a variável *Month*, começando pelas medições do mês de maio e finalizando com as do mês de setembro, escolhemos aleatoriamente 70% das observações para estimarmos os modelos e as outras 30% separamos para fins preditivos;

Passo 5: Maximizamos a função de log-verossimilhança da mesma forma que descrito no Passo 3 da Seção 4.2.2. Lembramos que, como estamos considerando quatro variáveis explicativas sem valores faltantes a mais do que as consideradas durante toda a construção do método, precisamos fazer os ajustes necessários para incorporá-las no modelo. Quanto às funções de log-verossimilhança maximizadas, temos que:

a) Para o método baseado em modelo proposto, a função de log-verossimilhança é a construída na Seção 4.3, em que consideramos os quatro cenários possíveis para cada observação do conjunto de treinamento, ou seja, a observação i ser observada em relação à variável X_1 (*Solar.R*) e ser faltante em relação à variável X_2 (*Ozone*); a observação i ser observada em relação à variável X_2 e ser faltante em relação à variável X_1; a observação i ser observada em relação às variáveis X_1 e X_2 ou a observação i ser faltante em relação às variáveis X_1 e X_2;

b) Para o método de imputação pela média, primeiramente calculamos a média dos valores do conjunto de treinamento correspondentes às variáveis X_1 e X_2 que não estão faltantes e, em seguida, para as observações que possuem valores faltantes em X_1 e X_2, imputamos as respectivas médias. Depois, com o conjunto de dados completo obtido com estas imputações, maximizamos a função de log-verossimilhança da

Seção 4.3 referente ao cenário em que as observações são completas em relação a X_1 e X_2;

c) Para o método de imputação por *Random Forest*, imputamos os valores faltantes em relação a X_1 e X_2 utilizando o pacote *missForest*. Após, com o conjunto de dados completo obtido com esta imputação, maximizamos a função de log-verossimilhança da Seção 4.3 referente ao cenário em que as observações são completas em relação a X_1 e X_2;

d) Para o método de imputação por *Hot-Deck*, imputamos os valores faltantes em relação a X_1 e X_2 utilizando o pacote *VIM*. Após, com o conjunto de dados completo, obtido com esta imputação, maximizamos a função de log-verossimilhança da Seção 4.3 referente ao cenário em que as observações são completas em relação a X_1 e X_2;

e) Para o método de imputação múltipla, imputamos os valores faltantes em relação a X_1 e X_2 utilizando o pacote *Amelia*. Por meio deste pacote, criamos cinco conjuntos de dados completos e, para encontrarmos a estimativa final dos parâmetros, calculamos a média aritmética das cinco estimativas para cada conjunto de dados completos (uma para cada conjunto de dados completos) encontradas por meio da maximização da função de log-verossimilhança da Seção 4.3 referente ao cenário em que as observações são completas em relação a X_1 e X_2;

Passo 6: Para analisarmos os métodos em relação ao poder preditivo, calculamos o erro quadrático do $\widehat{y}_i$ em relação ao y_i observado para todas as observações do conjunto de teste. O cálculo de $\widehat{y}_i$ se dá da seguinte forma:

a) Para a metodologia proposta e cada observação i do conjunto de dados de teste, calculamos $\widehat{y}_i$ de acordo com o proposto na Seção 4.3.1;

b) Para os métodos de imputação de dados por *Random Forest*, imputação por *Hot-Deck* e imputação múltipla, realizamos a imputação dos dados faltantes na base teste usando os mesmos procedimentos do passo 5 para cada método e, com os dados completos, calculamos $\widehat{y}_i$ de acordo com o item i) da Sessão 4.3.1. Para o caso da imputação múltipla, como criamos cinco conjuntos de dados completos, o erro quadrático é dado pela média entre os erros quadráticos dos cinco conjuntos gerados;

c) Para o método de imputação pela média, a média das observações não faltantes em X_1 do conjunto de treino é o valor utilizado para ser imputado nas observações que possuem valores faltantes no conjunto de teste para a variável X_1. Analogamente, a média das observações não faltantes em X_2 do conjunto de treino é o valor utilizado para ser imputado nas observações que possuem valores faltantes no conjunto de teste para a variável X_2. Após esta imputações, calculamos $\widehat{y}_i$ de acordo com o item i) da Sessão 4.3.1;

A Tabela 1 contém as estimativas para os parâmetros β_0, β_1, β_2, β_3, β_4, β_5, β_6, β_7 e σ_1^2 comuns para todos os métodos que utilizamos para fazer a aplicação nesses dados reais. Vale ressaltar que, para a implementação da metodologia baseada em modelo com resolução analítica, temos a estimação de mais 23 parâmetros, que não serão mostrados, já que são estimados apenas por este método.

Pela Tabela 1 podemos observar que o intercepto β_0 estimado é positivo para o método baseado em modelo e para a imputação múltipla e negativo para os demais métodos. As estimativas de β_4, associada à variável que representa o mês de maio são as mais distantes do zero em qualquer método e negativas, representando que esse mês apresenta, em média, as temperaturas mais baixas. Já as estimativas de β_5, β_6 e β_7, associadas a meses mais quentes que maio, quando o verão já se inicia em Nova York, são positivas. As estimativas de β_3, associada à covariável *Wind*, são todas negativas e o método baseado em modelo é o que estima a menor variância para o erro aleatório.

Tabela 1 – Estimativas dos parâmetros pelos métodos considerados. Aqui, MBM representa o método baseado em modelo, IMed a imputação pela média, IRF a imputação por *Random Forest*, IHD a imputação por *Hot-Deck* e IM a imputação múltipla.

Parâmetros	MBM	IMed	IRF	IHD	IM
β_0	0.099	-0.009	-0.063	-0.082	0.041
β_1	0.116	0.121	0.076	0.158	0.104
β_2	0.507	0.551	0.767	0.346	0.597
β_3	-0.148	-0.091	-0.033	-0.121	-0.054
β_4	-1.025	-0.943	-0.755	-1.026	-0.967
β_5	0.030	0.098	0.437	0.180	0.097
β_6	0.243	0.473	0.389	0.645	0.359
β_7	0.371	0.606	0.584	0.775	0.475
σ_1^2	0.188	0.244	0.211	0.316	0.207

Na Figura 85, temos as distâncias quadráticas entre valores observados e preditos no conjunto separado para teste. Foram detectados mais *outliers*, além dos já exibidos na figura. São eles:

- Método baseado em modelo com resolução analítica: 5.52;
- Método de imputação pela média: 6.81;
- Método de imputação por *random forest:* 10.38;
- Método de imputação por *hot-deck*: 6.64;
- Método de imputação múltipla: 6.89.

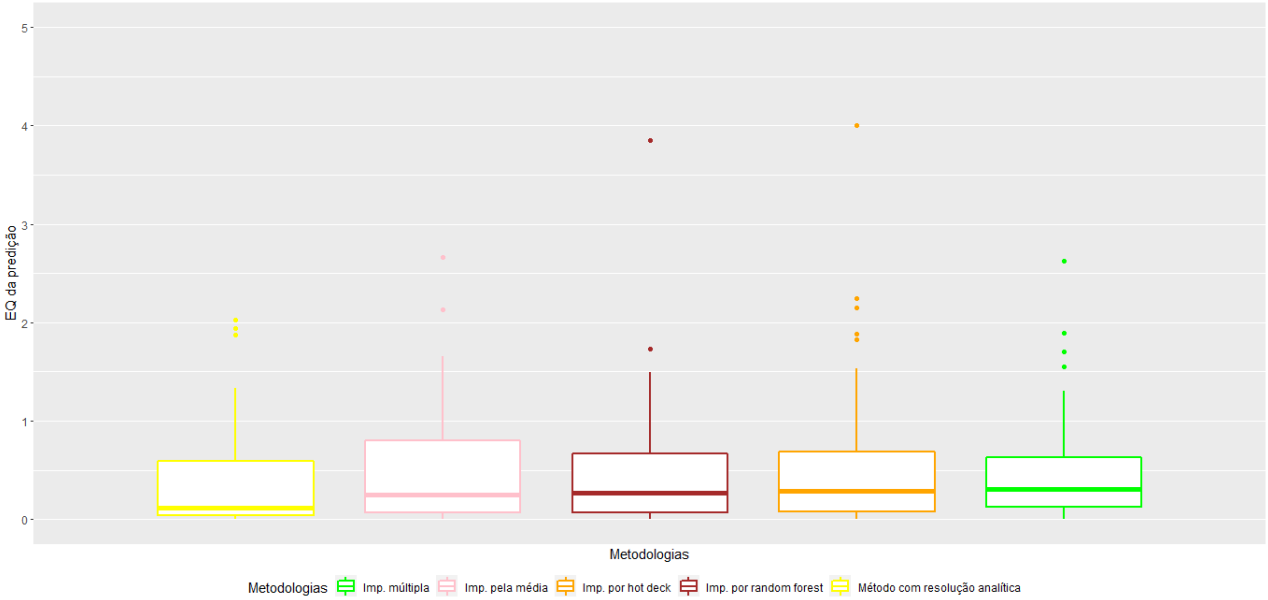

Figura 85 – Erro quadrático dos valores preditos para o conjunto de teste

Com podemos observar pela Figura 85, o método baseado em modelo com resolução analítica foi o que apresentou melhores resultados preditivos. A mediana e primeiro quartil das distâncias quadráticas entre valores observados e preditos são menores que de todos os outros métodos, além de possuir pouca variabilidade entre essas distâncias quadráticas.

Como todas as variáveis desse conjunto de dados assumem apenas valores positivos, quando não padronizadas, também faremos a análise usando a distribuição Weibull e os métodos de estimação sem resolução analítica.

6.3 Análise via modelo Weibull

Como todas as variáveis do conjunto de dados *Airquality* assumem valores positivos, realizamos a aplicação do método de integração por Monte Carlo utilizando o modelo Weibull para este *dataset* e compararmos os resultados com os demais métodos de imputação pela média, imputação por *Random Forest*, imputação por *Hot-deck* e imputação múltipla. Realizamos os seguintes procedimentos:

Passo 1: Consideramos *Temp* a variável dependente e *Solar.R, Ozone, Wind* e *Month* as variáveis explicativas;

Passo 2: Transformamos a variável categórica *Month*, que assume valores entre 5 e 9 de acordo com o mês a que se refere, em variáveis *dummies*, criando 4 colunas de variáveis dicotômicas, com valores 0 e 1 para os meses 5, 6, 7 e 8 e mantendo o mês 9 como categoria de referência;

Passo 3: Consideramos o log das variáveis *Solar.R* e *Ozone*, pois possuíam diferenças entre os valores máximo e mínimo consideráveis, e a raiz quadrada dos valores das variáveis *Temp* e *Wind*;

Passo 4: Como os dados são ordenados de acordo com a variável *Month*, começando pelas medições do mês de maio e finalizando com as do mês de setembro, escolhemos aleatoriamente 70% das observações para estimarmos os modelos e as outras 30% separamos para fins preditivos;

Passo 5: Adaptamos as implementações dos passos 4 ao 7 da Seção 5.2.5 para esta aplicação a um conjunto de dados reais.

A Tabela 2 contém as estimativas para os parâmetros β_0, β_1, β_2, β_3, β_4, β_5, β_6, β_7 e α_1 comuns para todos os métodos que utilizamos para fazer a aplicação nesses dados reais.

Pela Tabela 2 podemos observar que o intercepto β_0 estimado é positivo para todos os métodos considerados. As estimativas de β_4, associada à variável que representa o mês de maio são as mais distantes do zero para o método baseado no modelo Weibull, representando que esse mês apresenta, em média, as temperaturas mais baixas. Já as estimativas de β_5, β_6 e β_7, associadas a meses mais quentes que maio, quando o verão já se inicia em Nova York, vão aumentando gradualmente.

Tabela 2 – Estimativas dos parâmetros pelos métodos considerados. Aqui, MBM representa o método baseado em modelo, IMed a imputação pela média, IRF a imputação por *Random Forest*, IHD a imputação por *Hot-Deck* e IM a imputação múltipla.

Parâmetros	MBM	IMed	IRF	IHD	IM
β_0	1.964	0.911	1.064	1.925	1.420
β_1	0.016	0.019	-0.057	0.085	0.107
β_2	0.033	0.158	0.251	-0.053	0.020
β_3	0.011	0.208	0.174	0.017	0.072
β_4	-0.053	-0.072	0.095	-0.045	-0.072
β_5	-0.008	0.107	0.027	-0.027	-0.003
β_6	0.015	0.007	-0.005	0.044	0.012
β_7	0.029	0.033	-0.014	0.082	0.021
α_1	29.234	6.291	6.647	9.585	6.600

Na Figura 86, temos as distâncias quadráticas entre valores observados e preditos no conjunto separado para teste. Foram detectados mais *outliers*, além dos já exibidos na figura. São eles:

- Método de integração por Monte Carlo: 6.0319×10^5, 6.3009×10^5, 6.3215×10^5, 6.4391×10^5, 6.2094×10^5 e 6.1954×10^5;
- Método de imputação pela média: 11.58, 20.77, 3.92 e 3.58;
- Método de imputação por *random forest:* 9.84, 6.37, 5.53 e 4.54;
- Método de imputação por *hot-deck*: 3.37;
- Método de imputação múltipla: 4.67 e 5.28.

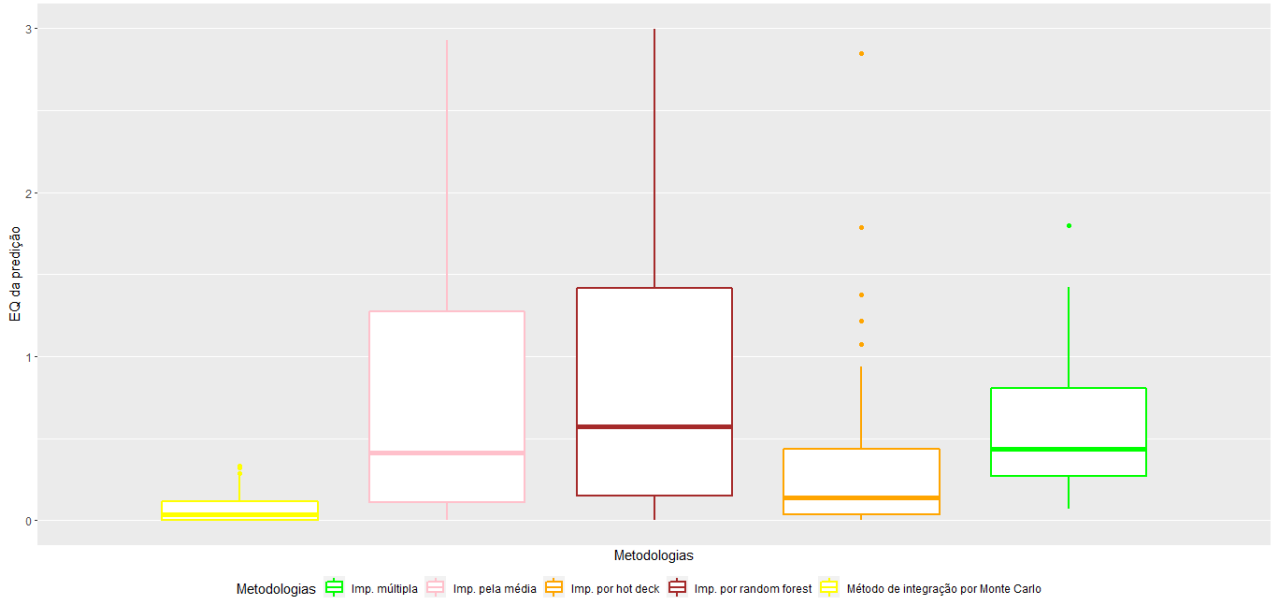

Figura 86 – Erro quadrático dos valores preditos para o conjunto de teste

Com podemos observar pela Figura 86, o método baseado em modelo sem resolução analítica utilizando o método de integração por Monte Carlo foi o que apresentou melhores resultados preditivos. A mediana e primeiro quartil das distâncias quadráticas entre valores observados e preditos são menores que de todos os outros métodos, além de possuir pouca variabilidade entre essas distâncias quadráticas.

Neste capítulo, realizamos a aplicação de dois métodos no conjunto de dados *Airquality*, uma via modelo Gaussiano e outra via modelo Weibull. Como uma forma de definir qual dos dois modelos melhor se ajusta a estes dados, pode-se desenvolver uma análise de resíduos, em que seria verificado se os modelos propostos capturam toda a estrutura de dependência entre a variável modelada e as explicativas. Critérios de seleção de modelos como o AIC ou BIC também poderia ser utilizado.

CAPÍTULO

7

CONCLUSÃO

Neste trabalho introduzimos os principais conceitos relacionados aos dados faltantes, incluindo três características importantes relacionadas a eles que auxiliam na escolha de um método adequado para o tratamento dos mesmos. A primeira dessas características é o mecanismo associado aos dados faltantes, o qual está relacionado com a dependência entre valores de diferentes atributos, podendo ser: independente de qualquer evento (MCAR), dependente de outros atributos (MAR) ou dependente do próprio atributo (MNAR). A segunda dessas características é o padrão. O padrão indica se os dados faltantes ocorrem de forma não-estruturada ou sistemática. A terceira e última dessas características é a quantidade de dados faltantes (VERONEZE, 2011).

Em seguida, foram descritos alguns dos métodos mais conhecidos para o tratamento de dados faltantes, como também as vantagens e desvantagens de cada um desses métodos. Eles são divididos em baseados na deleção (*listwise*, *pairwise* e variáveis descartáveis) e imputação de dados (métodos específicos de séries temporais, imputação pela média, mediana e moda, imputação múltipla, imputação por *Random forest*, imputação por *Hot-deck*, imputação de variáveis categóricas e *K Nearest Neighbors*). Quanto aos métodos de deleção de casos, os resultados obtidos se tornam viesados se os casos restantes não representarem toda a população. Além disso, a não inclusão no modelo das variáveis que possuem dados faltantes, embora não acarrete problemas de enviesamento da base de estimação, pode ser determinante na obtenção de um modelo com um poder preditivo inferior ao que seria resultado caso todas as variáveis fossem testadas. Já em relação aos métodos de imputação de dados, deve-se levar em consideração que os valores imputados não são valores reais, logo, é essencial trabalhar com a incerteza associada à imputação, visando validar os resultados obtidos com os dados completos.

Com base em todos essas questões, propomos nesse trabalho algumas metodologias baseadas em modelos que não envolvem o processo de remoção de dados e nem o de imputação. Com isso, não precisamos nos importar com as caraterísticas dos dados faltantes, tais como o mecanismo gerador dos mesmos, o padrão ou a quantidade dos valores faltantes, além das

metodologias se ajustarem a quaisquer modelos (lineares e não lineares) que envolvam variáveis resposta e independentes.

Para verificarmos a eficácia dos métodos propostos, realizamos estudos de simulação considerando diferentes tamanhos de amostra, os três tipos diferentes de mecanismos de geração de valores faltantes (MCAR, MAR e MNAR), o número de variáveis com valores faltantes nos conjunto de dados (uma variável ou duas) e a proporção de valores faltantes em cada variável (20% ou 60%).

Considerando o modelo Gaussiano, para o qual a metodologia apresenta resultado exato e analítico, com uma e duas variáveis faltantes, observamos resultados inferenciais e preditivos satisfatórios em relação aos métodos comparados (deleção de observações, imputação pela média, imputação por *random forest*, imputação por *hot-deck*, imputação múltipla e dados completos). Em ambas as configurações referentes ao número de variáveis com valores faltantes, para o mecanismo MNAR de geração dos dados faltantes, que provavelmente é a situação de estimação e predição mais desafiadora, o método proposto apresenta, de maneira geral, um desempenho de estimação superior a todos os métodos comparados, exceto à estimação realizada com os dados completos. Esse comportamento acontece mesmo o método apresentado tendo que estimar 8 parâmetros adicionais, para o caso com duas variáveis com valores faltantes. O método de estimação baseado no modelo Gaussiano também apresenta resultados preditivos superiores a todos os métodos, exceto ao modelo estimado com dados completos. Também foi realizada uma aplicação para dados reais climáticos e de qualidade do ar com o método baseado em modelo com resolução analítica comparativamente as metodologias utilizadas nos estudos de simulação, em que o método MMORA apresenta resultados muito satisfatórios e melhores que os dos outros métodos em relação às distâncias quadráticas dos valores preditos para o conjunto de teste.

Em relação aos métodos baseados em modelo sem resolução analítica, realizamos simulações considerando a distribuição Weibull para as variáveis resposta e independentes, mas vale ressaltar que os métodos se adequam a qualquer outro modelo. Para o caso com apenas uma variável com valores faltantes, utilizando o método de integração por Monte Carlo e considerando o mecanismo MCAR (mecanismo ignorável), obtivemos resultados inferenciais muito parecidos à análise dos dados completos, especialmente para amostras maiores ($n = 300$), em que ele se sobressai inclusive ao método de imputação múltipla, considerado o método padrão para lidar com dados faltantes até então. Considerando o mecanismo MAR (mecanismo ignorável), observamos resultados inferenciais e preditivos muito satisfatórios, especialmente em relação ao viés das estimativas dos parâmetros, em que ele se equipara ao método de imputação múltipla e, para amostras maiores ($n = 300$), consegue ter melhor performance que a dele. Considerando o mecanismo MNAR (não-ignorável), para amostras maiores, o método por integração de Monte Carlo apresenta ótimos resultados inferenciais, especialmente se observarmos o erro quadrático médio das estimativas dos parâmetros. Além disso, para amostras menores, ele apresenta resultados preditivos melhores do que metodologias que são focadas na predição, como

o método de imputação múltipla.

Para o caso com apenas uma variável com valores faltantes, considerando o mecanismo MCAR e amostras menores ($n = 100$), apesar do viés das estimativas do método de média de log-verossimilhanças ser pior que o do método da integração por Monte Carlo, ele apresenta melhores resultados em relação ao erro quadrático médio. Isso provavelmente devido ao fato de que a variância dos estimadores por esse método de média de log-verossimilhanças é menor que a variância dos estimadores por integração de Monte Carlo e isso compensou o viés no cálculo do EQM. Considerando o mecanismo MAR, este método apresenta resultados inferenciais muito parecidos com a estimação considerando os dados completos e o método de imputação múltipla. Para os cenários com maior proporção de valores faltantes ($p = 60\%$), ele se sobressai em relação ao método de imputação múltipla, tanto em relação ao viés e erro quadrático médio das estimativas dos parâmetros, quanto em relação à performance preditiva. Considerando o mecanismo MNAR, este método, juntamente com o utilizando integração por Monte Carlo e os métodos de imputação múltipla, deleção de casos e dados completos foram os que apresentaram melhores resultados em relação à estimação dos parâmetros. Para amostras menores ($n = 100$), levando-se em consideração o erro quadrático médio das estimativas dos parâmetros, o método utilizando média de log-verossimilhanças e o com dados completos se destacam em relação a todos os outros métodos.

Para o caso com uma variável com valores faltantes, de modo geral, o método utilizando o algoritmo EM obteve resultados inferenciais ruins, o que se deve ao fato desse algoritmo ser sensível aos valores iniciais. Como estamos tomando valores iniciais não informativos, podemos ter casos em que o método converge para máximo local e não global ou, às vezes, nem converge. Uma das maneiras de melhorarmos a estimação e predição pelo método utilizando o algoritmo EM seria iniciarmos o método através de valores iniciais mais informativos, provavelmente calculados através dos dados disponíveis. Aqui, para justa comparação entre os diferentes métodos, não fizemos isso.

Como, de modo geral, para uma variável com valores faltantes, dentre os métodos propostos baseados em modelos sem resolução analítica, o método de integração por Monte Carlo foi o que apresentou melhores resultados, além de possuir menor custo computacional e, como ele já obteve resultados satisfatórios com amostras menores ($n = 100$), desenvolvemos estudos de simulação para duas variáveis com valores faltantes apenas para este método em comparativo com os métodos de deleção de observações, imputação pela média, imputação por *random forest*, imputação por *hot-deck*, imputação múltipla e dados completos, com amostras de tamanho 100 e para os cenários MCAR e MNAR, pois os resultados apresentados nos estudos de simulação anteriores para os cenários MCAR e MAR foram parecidos. Nestes estudos, observamos que, para o cenário MCAR com maior proporção de valores faltantes na variável X_1, o método de integração por Monte Carlo, juntamente com os métodos de imputação múltipla e método considerando os dados completos foram os que obtiveram melhores resultados

inferenciais, sendo que o método proposto ultrapassou o método de imputação múltipla quanto ao poder preditivo. Para o cenário MNAR, observamos que novamente estes três métodos, método de integração por Monte Carlo, método de imputação múltipla e método considerando os dados completos, obtiveram as melhores performances inferenciais, enquanto que, em relação ao poder preditivo, embora a variabilidade do método de imputação múltipla para os erros quadráticos médios dos valores preditos no conjunto de teste seja a menor, ele é o que apresenta mais e maiores *outliers*, além do fato de a mediana do método baseado em modelo e do método com dados completos serem menores que a do método de imputação múltipla, para o caso com maior proporção de valores faltantes na variável X_1. Vale ressaltar que o método de imputação múltipla, apesar de mostrar bons resultados no geral, se mostra um dos métodos mais instáveis em termos de resultados, pois em alguns cenários mostra um comportamento ótimo e em outros um comportamento ruim. Além de mostrar lentidão de processamento em cenários mais desafiadores (com mais proporção de *missings* e mecanismo MNAR de dados faltantes). Também realizamos a aplicação do método baseado na integração de Monte Carlo para o conjunto de dados *Airquality*, em que observamos a satisfatória performance preditiva deste método em relação aos demais métodos utilizados para comparação.

A principal desvantagem dos métodos desenvolvidos baseados em modelos é que, para o caso sem resolução analítica, o tempo para estimação do modelo é maior que o dos outros métodos que usamos para comparação, especialmente para amostras maiores, com proporção de dados faltantes de 60% e sob o mecanismo MNAR. Comparando o tempo computacional entre os três métodos baseados em modelos sem resolução analítica (utilizando integração por Monte Carlo, média de log-verossimilhanças e algoritmo EM), o método de integração por Monte Carlo é o mais vantajoso e o do algoritmo EM o mais custoso computacionalmente. No entanto, vale ressaltar que os resultados aqui mostrados são baseados em 500 (e em poucos casos 50) valores simulados para cada dado faltante em cada iteração dos algoritmos. Se reduzirmos esse valor, o tempo de processamento pode cair drasticamente sem mudar significativamente os resultados. Como estudos futuros, podemos realizar uma análise de sensibilidade dos métodos a esse valor e indicar um número mínimo a ser utilizado, sem prejudicar a qualidade dos resultados.

Enfim, como propostas futuras, para situações que contemplem mais de duas variáveis faltantes, os métodos aqui apresentados podem ser estendidos com base na generalização que fizemos do caso com uma variável faltante para duas variáveis faltantes. Nesses casos, os métodos propostos também podem ser consecutivamente utilizados em algumas etapas. Primeiramente, eles podem ser aplicados para imputar dados em algumas variáveis faltantes e, quando restar apenas um ou duas variáveis com dados faltantes, serem utilizados para estimar o modelo final.

REFERÊNCIAS

ACUNA, E.; RODRIGUEZ, C. The treatment of missing values and its effect on classifier accuracy. In: **Classification, clustering, and data mining applications**. [S.l.]: Springer, 2004. p. 639–647. Citado na página 38.

ALLISON, P. D. Estimation of linear models with incomplete data. **Sociological Methodology**, JSTOR, p. 71–103, 1987. Citado na página 23.

____. **Missing data**. [S.l.]: Sage publications, 2001. Citado nas páginas 33, 34, 35, 39, 40, 41, 43, 44 e 45.

ASSUNÇÃO, F. **Estratégias para tratamento de variáveis com dados faltantes durante o desenvolvimento de modelos preditivos**. Tese (Doutorado) — Universidade de São Paulo, 2012. Citado nas páginas 22, 23 e 36.

BLACKWELL, M.; HONAKER, J.; KING, G. A unified approach to measurement error and missing data: overview and applications. **Sociological Methods & Research**, SAGE Publications Sage CA: Los Angeles, CA, v. 46, n. 3, p. 303–341, 2017. Citado na página 24.

BROWN, M. L.; KROS, J. F. Data mining and the impact of missing data. **Industrial Management & Data Systems**, MCB UP Ltd, v. 103, n. 8, p. 611–621, 2003. Citado nas páginas 21, 37 e 38.

BUUREN, S. van; MULLIGEN, E. van; BRAND, J. P. Routine multiple imputation in statistical databases. In: IEEE. **Seventh International Working Conference on Scientific and Statistical Database Management**. [S.l.], 1994. p. 74–78. Citado nas páginas 21 e 22.

CHAMBERS, J. M.; CLEVELAND, W. S.; KLEINER, B.; TUKEY, P. A. **Graphical methods for data analysis**. [S.l.]: Chapman and Hall/CRC, 2018. Citado na página 159.

CHE, Z.; PURUSHOTHAM, S.; CHO, K.; SONTAG, D.; LIU, Y. Recurrent neural networks for multivariate time series with missing values. **Scientific Reports**, Nature Publishing Group, v. 8, n. 1, p. 1–12, 2018. Citado na página 24.

COLANTONIO, A.; PIETRO, R. D.; OCELLO, A.; VERDE, N. V. Abba: Adaptive bicluster-based approach to impute missing values in binary matrices. In: **Proceedings of the 2010 ACM Symposium on Applied Computing**. [S.l.: s.n.], 2010. p. 1026–1033. Citado nas páginas 21 e 22.

FARHANGFAR, A.; KURGAN, L. A.; PEDRYCZ, W. Experimental analysis of methods for imputation of missing values in databases. In: SPIE. **Intelligent Computing: Theory and Applications II**. [S.l.], 2004. v. 5421, p. 172–182. Citado nas páginas 21, 22, 37 e 38.

____. A novel framework for imputation of missing values in databases. **IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans**, IEEE, v. 37, n. 5, p. 692–709, 2007. Citado nas páginas 21, 22 e 38.

FICHMAN, M.; CUMMINGS, J. N. Multiple imputation for missing data: Making the most of what you know. **Organizational Research Methods**, Sage Publications, v. 6, n. 3, p. 282–308, 2003. Citado nas páginas 33 e 34.

GRAHAM, J. W.; CUMSILLE, P. E.; ELEK-FISK, E. Methods for handling missing data. John Wiley & Sons Inc, 2003. Citado na página 28.

GRAHAM, J. W.; HOFER, S. M.; DONALDSON, S. I.; MACKINNON, D. P.; SCHAFER, J. L. Analysis with missing data in prevention research. American Psychological Association, 1997. Citado nas páginas 28 e 29.

GRAHAM, J. W.; HOFER, S. M.; PICCININ, A. M. Analysis with missing data in drug prevention research. **NIDA Research Monograph**, Citeseer, v. 142, p. 13–13, 1994. Citado na página 34.

GRAHAM, J. W. *et al.* Missing data analysis: Making it work in the real world. **Annual Review of Psychology**, Palo Alto, v. 60, n. 1, p. 549–576, 2009. Citado nas páginas 22, 29 e 42.

HONAKER, J.; KING, G.; BLACKWELL, M. **A Program for Missing Data**. [S.l.], 2021. R package version 1.8.0. Disponível em: <https://cran.r-project.org/web/packages/Amelia/Amelia.pdf>. Citado na página 58.

JR, E. R. H.; EBECKEN, N. F. Missing values prediction with k2. **Intelligent Data Analysis**, IOS Press, v. 6, n. 6, p. 557–566, 2002. Citado na página 45.

KIM, J.-O.; CURRY, J. The treatment of missing data in multivariate analysis. **Sociological Methods & Research**, Sage Publications Sage CA: Thousand Oaks, CA, v. 6, n. 2, p. 215–240, 1977. Citado na página 35.

KURASOVA, O.; MARCINKEVICIUS, V.; MEDVEDEV, V.; RAPECKA, A.; STEFANOVIC, P. Strategies for big data clustering. In: IEEE. **2014 IEEE 26th international conference on tools with artificial intelligence**. [S.l.], 2014. p. 740–747. Citado na página 21.

LAKSHMINARAYAN, K.; HARP, S. A.; SAMAD, T. Imputation of missing data in industrial databases. **Applied Intelligence**, Springer, v. 11, n. 3, p. 259–275, 1999. Citado nas páginas 21, 37, 38, 43 e 45.

LITTLE, R. J. A test of missing completely at random for multivariate data with missing values. **Journal of the American Statistical Association**, Taylor & Francis, v. 83, n. 404, p. 1198–1202, 1988. Citado na página 29.

LOPES, M. M. **Programação Genética para otimização de séries temporais com dados faltantes**. Tese (Doutorado) — Universidade Federal do Rio de Janeiro, 2007. Citado na página 22.

LUENGO, J.; GARCÍA, S.; HERRERA, F. On the choice of the best imputation methods for missing values considering three groups of classification methods. **Knowledge and Information Systems**, Springer, v. 32, n. 1, p. 77–108, 2012. Citado na página 45.

MANLY, C. A.; WELLS, R. S. Reporting the use of multiple imputation for missing data in higher education research. **Research in Higher Education**, Springer, v. 56, n. 4, p. 397–409, 2015. Citado na página 24.

MARLIN, B. **Missing data problems in machine learning**. Tese (Doutorado) — University of Toronto, 2008. Citado na página 22.

MCKNIGHT, P. E.; MCKNIGHT, K. M.; SIDANI, S.; FIGUEREDO, A. J. **Missing data: A gentle introduction**. [S.l.]: Guilford Press, 2007. Citado nas páginas 27, 29, 30, 31, 35, 36, 37, 38, 39, 41, 42, 43, 44 e 45.

MENG, X.-L.; RUBIN, D. B. Maximum likelihood estimation via the ECM algorithm: A general framework. **Biometrika**, Oxford University Press, v. 80, n. 2, p. 267–278, 1993. Citado na página 42.

MUTHÉN, B.; KAPLAN, D.; HOLLIS, M. On structural equation modeling with data that are not missing completely at random. **Psychometrika**, Springer, v. 52, n. 3, p. 431–462, 1987. Citado na página 23.

MYRTVEIT, I.; STENSRUD, E.; OLSSON, U. H. Analyzing data sets with missing data: An empirical evaluation of imputation methods and likelihood-based methods. **IEEE Transactions on Software Engineering**, IEEE, v. 27, n. 11, p. 999–1013, 2001. Citado nas páginas 22, 27, 37 e 40.

NUNES, L. N.; KLÜCK, M. M.; FACHEL, J. M. G. Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos. **Cadernos de Saúde Pública**, SciELO Brasil, v. 25, p. 268–278, 2009. Citado nas páginas 22 e 23.

PAES, A.; POLETO, F. Z. O problema de dados omissos (missing data). **Educação Continuada em Saúde: Einsten**, v. 11, n. 1, p. 5–7, 2013. Citado na página 29.

PEREIRA, E. A. **Algumas propostas para imputaçao de dados faltantes em teoria de resposta ao item**. Dissertação (Mestrado) — Universidade de Brasília, 2014. Citado nas páginas 22 e 27.

PIMENTEL-ALARCÓN, D.; BALZANO, L.; MARCIA, R.; NOWAK, R.; WILLETT, R. Group-sparse subspace clustering with missing data. In: IEEE. **2016 IEEE Statistical Signal Processing Workshop (SSP)**. [S.l.], 2016. p. 1–5. Citado na página 24.

POLETO, F. Z. **Analise de dados categorizados com omissao**. Dissertação (Mestrado) — Universidade de São Paulo, 2006. Citado na página 46.

POLETO, F. Z. **Análise de dados categorizados com omissão em variáveis explicativas e respostas**. Tese (Doutorado) — Universidade de Sao Paulo, 2011. Citado na página 24.

PRASS, F. S. **Estudo comparativo entre algoritmos de análise de agrupamentos em data mining**. Dissertação (Mestrado) — Centro Tecnológico, Universidade Federal de Santa Catarina, 2014. Citado na página 21.

RIBEIRO, E. A. **Imputação de dados faltantes via algoritmo EM e rede neural MLP com o método de estimativa de máxima verossimilhança para aumentar a acurácia das estimativas**. Dissertação (Mestrado) — Universidade Federal de Sergipe, 2015. Citado nas páginas 21 e 34.

RUBIN, D. B. **Multiple Imputation for Nonresponse in Surveys**. [S.l.]: John Wiley & Sons, 1987. Citado nas páginas 22 e 42.

RUBIN, D. B.; LITTLE, R. J. **Statistical analysis with missing data**. [S.l.]: John Wiley & Sons, 2019. Citado nas páginas 22, 28, 30, 34, 35, 39, 40, 41, 42 e 44.

SADINLE, M.; REITER, J. P. Sequentially additive nonignorable missing data modelling using auxiliary marginal information. **Biometrika**, Oxford University Press, v. 106, n. 4, p. 889–911, 2019. Citado na página 24.

SANTANA, I. F.; FILIZOLA-NETO, N. P.; FREITAS, C. E. de C. Recuperação de valores perdidos "missing value" de dados de desembarque: Uma aplicação com dados de desembarque de semaprochilodus sp. em Santarém, Estado do Pará, Brasil. **Revista Brasileira de Engenharia de Pesca**, v. 5, n. 1, p. 43–55, 2010. Citado na página 22.

SCHAFER, J. L. **Analysis of incomplete multivariate data**. [S.l.]: CRC press, 1997. Citado nas páginas 36 e 43.

SCHAFER, J. L.; GRAHAM, J. W. Missing data: our view of the state of the art. **Psychological Methods**, American Psychological Association, v. 7, n. 2, p. 147, 2002. Citado nas páginas 29, 35, 37, 42, 43 e 45.

SORJAMAA, A. *et al.* Methodologies for time series prediction and missing value imputation. Multiprint, 2010. Citado na página 28.

STEKHOVEN, D. J. **Using the missForest Package**. [S.l.], 2011. R package version 1.5. Disponível em: <https://cran.r-project.org/web/packages/missForest/vignettes/missForest_1.5.pdf>. Citado na página 57.

TAN, M.; TSANG, I. W.; WANG, L. Towards ultrahigh dimensional feature selection for big data. **Journal of Machine Learning Research**, 2014. Citado na página 21.

TANNER, M. A.; WONG, W. H. The calculation of posterior distributions by data augmentation. **Journal of the American Statistical Association**, Taylor & Francis, v. 82, n. 398, p. 528–540, 1987. Citado na página 23.

TEMPL, M.; KOWARIK, A.; ALFONS, A.; CILLIA, G. de; PRANTNER, B.; RANNETBAUER, W. **Visualization and Imputation of Missing Values**. [S.l.], 2021. R package version 6.1.1. Disponível em: <https://cran.r-project.org/web/packages/VIM/VIM.pdf>. Citado na página 58.

VERONEZE, R. **Tratamento de dados faltantes empregando biclusterização com imputação múltipla**. Tese (Doutorado) — Universidade Estadual de Campinas, 2011. Citado nas páginas 27, 28, 31, 35, 36, 39, 45 e 171.

WANG, Y.; LI, B.; LUO, R.; CHEN, Y.; XU, N.; YANG, H. Energy efficient neural networks for big data analytics. In: IEEE. **2014 Design, Automation & Test in Europe Conference & Exhibition (DATE)**. [S.l.], 2014. p. 1–2. Citado na página 21.

WU, C.-H.; WUN, C.-H.; CHOU, H.-J. Using association rules for completing missing data. In: IEEE. **Fourth International Conference on Hybrid Intelligent Systems (HIS'04)**. [S.l.], 2004. p. 236–241. Citado nas páginas 21, 22 e 27.

ZHANG, P. Multiple imputation: theory and method. **International Statistical Review/Revue Internationale de Statistique**, JSTOR, p. 581–592, 2003. Citado nas páginas 28, 29 e 43.


ICMC
USP
SÃO CARLOS