Universidade de Brasília
Departamento de Estatística

Técnicas de imputação de dados faltantes com uso de modelagem preditiva

Enzo Porto Brasil

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

Brasília
2023

Enzo Porto Brasil

Técnicas de imputação de dados faltantes com uso de modelagem preditiva

Orientador: Prof. Eduardo Yoshio Nakano

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

Brasília
2023

Agradecimentos

Gostaria de expressar minha profunda gratidão por trás de um trabalho com tamanha dedicação. Agradeço aos meus pais, Carlos e Valquiria, que incentivaram cada passo em minha vida, com extremo apoio aos meus estudos, com amor representado pela liberdade e preparação para todos as situações de minha vida. Aos meus irmãos Déric, João, Luma e Maria, que foram luz e força na minha jornada. Aos meus avós Brasil, Lúcia e Socorro, que foram exemplos de carinho.

À minha namorada Jéssica, que, com muito empenho e ternura, foi paciente e participou de toda a minha vida acadêmica, e esteve nos melhores momentos da minha vida. Agradeço ao amor de toda a minha família Porto e família Brasil, que em todos esses anos viram minha evolução e não exitaram em participar dela. À família Weschenfelder Ferreira, que encorajaram ainda mais minhas realizações.

A todos os meus colegas e amigos da graduação na Universidade de Brasília, os quais contribuíram com todas as forças para meus estudos e desempenho, com grupos de estudo e construção de ideias. Aos amigos próximos, que mesmo em momentos e locais distantes, incentivaram cada passo da minha jornada.

Aos professores incríveis que tive durante todo meu ensino, desde o ensino médio ao superior. Com gratidão especial ao professor orientador Nakano, o qual foi compreensível e atencioso em todas as nossas pesquisas na Universidade, agradeço por acreditar no meu potencial e ser um mentor incomparável. Também aos professores Leandro, José e André, que participaram do desenvolvimento e aperfeiçoamento deste trabalho.

Resumo

A presença de dados faltantes em um banco de dados é comum, mas as análises estatísticas são fundamentadas para analisar dados completos. Uma das possíveis soluções para o tratamento da ausência de informações é utilizar imputação de dados. Este trabalho teve o objetivo de destacar o desempenho e importância de técnicas de imputação que utilizam regressão estatística, e propor assim um guia eficaz e prático aos pesquisadores com base na literatura e nos resultados observados. Além do estudo dos dois pilares principais desta pesquisa, regressão estatística e técnicas de imputação, a metodologia implementada foi feita por meio de simulações, com a criação de bancos de dados artificiais, possibilitando a discussão de diferentes cenários de dados faltantes e técnicas. Questões relacionadas ao impacto da quantidade de dados faltantes, da natureza da variável incompleta e da quantidade de variáveis incompletas também foram verificadas. A partir de ajustes de modelos de regressão linear, os tipos de imputação (única e múltipla) e técnicas de imputação (medidas de tendência central e modelagem preditiva) foram comparados. Os resultados obtidos, junto às discussões, foram favoráveis ao uso de modelos de regressão em processos de imputação de dados, com evidências da importância ao selecionar métodos adequados com base nos motivos da ausência de dados e na proporção de dados faltantes.

Palavras-chave: Dados faltantes; Imputação única; Imputação múltipla; Técnicas de imputação; Regressão estatística; Modelagem preditiva; Simulação; Guia para imputação.

Abstract

The presence of missing data in a database is common, but statistical analyses are typically performed on complete data. One possible solution for handling missing information is to use data imputation. This work aimed to highlight the performance and importance of imputation techniques that use statistical regression, and to propose an effective and practical guide for researchers based both on literature and the observed results. In addition to studying the two main pillars of this research, statistical regression and imputation techniques, the methodology implemented was done through simulations, with the creation of artificial databases, allowing for the discussion of different scenarios of missing data and techniques. Issues related to the impact of the amount of missing data, the nature of the incomplete variable, and the amount of incomplete variables were also verified. Based on adjustments of linear regression models, the types of imputation (single and multiple) and imputation techniques (measures of central tendency and predictive modelling) were compared. The results obtained, along with the discussions, were favorable to the use of regression models in data imputation processes, with evidence of the importance of selecting appropriate methods based on the reasons for missing data and the proportion of missing data.

Keywords: Missing data; Single imputation; Multiple imputation; Imputation techniques; Statistical regression; Predictive modelling; Simulation; Guide for imputation.

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

Os dados se tornaram um dos objetos de estudo mais valiosos da ciência no século 21. Por consequência, apesar de facilitado, o acesso à grande quantidade de informações possibilita investigações capazes de gerar resultados tendenciosos, os quais interferem de forma considerável na propagação de *fake news* e conclusões inadequadas (PROVOST; FAWCETT, 2013). Além disso, muito material é perdido durante um processo de pesquisa, e desconsiderar tais perdas de observações pode não ser a forma mais adequada de avaliar determinados fenômenos (NUNES, 2007).

Informações ausentes em um banco de dados é muito recorrente, e a devida atenção a isso pode modificar por completo a forma que os resultados são interpretados (RUBIN, 1996). Como os métodos estatísticos utilizados em estudos são fundamentados para analisar dados completos, o tratamento adequado do tema não deve ser ignorado (HARRELL *et al.*, 2015).

Neste contexto, a imputação de dados faltantes é fundamental para a elaboração de processos de preenchimento de valores ausentes em um conjunto de dados. A imputação pode ser feita baseada em procedimentos simples, como utilizar estatísticas descritivas ou substituir por valores de outra variável semelhante à que apresenta dados faltantes. Porém, em muitas situações, o uso de modelos estatísticos sofisticados para imputar tais valores é capaz de transformar positivamente a leitura de resultados (MCKNIGHT *et al.*, 2007; NUNES, 2007).

Além disso, a depender do tipo de assunto a ser estudado, a ocorrência de dados faltantes requer maior atenção. Por exemplo, na área de saúde pública, saber a quantidade de pessoas que desenvolveram determinada doença psicológica ao longo do tempo em certo local pode interferir diretamente nas medidas adotadas para melhorar a situação dessa região, e ausência desses dados dificulta na elaboração de ações efetivas para solucionar o problema (HEIJDEN *et al.*, 2006).

O estudo sobre imputação de dados faltantes não está relacionado apenas com a teoria estatística, mas também com a maneira que diversas situações podem ser reavaliadas e prevenidas, de acordo com cada tema a ser pesquisado. Dessa forma, o impacto gerado pela recuperação de dados perdidos, sejam eles informações ausentes por fator aleatório ou não, pode interferir de maneira significativa nas possíveis decisões a serem feitas (KENWARD; CARPENTER, 2007; BUUREN, 2018).

Portanto, este trabalho teve sobretudo o objetivo de analisar e avaliar as principais técnicas de imputação de dados faltantes com base em modelagem estatística, e relatar a importância de tais métodos para diferentes cenários no meio científico. A escolha do modelo adequado para cada imputação ocorreu de acordo com a sua capacidade preditiva

e a natureza da variável a ser imputada. Foi requisito para o estudo analisar apenas bases de dados compostas por dados faltantes gerados por fator aleatório. Também foram explorados conceitos pouco divulgados relacionados à área de imputação, com elaboração de um guia prático e possíveis sugestões para futuras pesquisas relacionadas ao tema. Além do mais a metodologia apresentada neste trabalho foi ilustrada por dados simulados, e o desenvolvimento das técnicas ocorreu com auxílio do software R/Rstudio versão 4.3.2 (R Core Team, 2023).

2 Regressão Estatística

A análise de regressão é uma metodologia estatística que se baseia no estudo da relação entre duas ou mais variáveis, de modo que uma variável, caracterizada como resposta, possa ser explicada a partir de uma ou mais variáveis explicativas. A natureza da variável resposta é o que define quais técnicas de modelagem devem ser consideradas para avaliar a sua relação estatística com as variáveis explicativas (DRAPER; SMITH, 1998; KUTNER *et al.*, 2005).

2.1 Regressão Linear

A Regressão Linear trata do estudo da relação estatística entre duas ou mais variáveis, de modo que a variável resposta tenha natureza quantitativa discreta ou contínua. Essa metodologia pode ser classificada como regressão linear simples ou múltipla (RODRIGUES, 2012).

Quanto às condições de linearidade, Kutner *et al.* (2005) explicam:

> [...] *estes modelos são considerados lineares nos parâmetros porque nenhum parâmetro aparece como um expoente ou, é multiplicado ou dividido por outro parâmetro. E lineares na variável explicativa, porque essa variável aparece apenas elevada a um (equação de primeiro grau).*

2.1.1 Regressão Linear Simples

A Regressão Linear Simples é estruturada com base na análise da relação entre apenas duas variáveis, sendo uma caracterizada como explicativa (também chamada de variável independente ou preditora; a variável que explica o fenômeno) e outra como variável resposta (variável que comporta os dados de interesse a serem obtidos). Dessa forma, o modelo linear simples é comumente denotado como (KUTNER *et al.*, 2005; HARRELL *et al.*, 2015)

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \tag{2.1.1}$$

em que:

- Y_i é o valor da variável resposta ou dependente na i-ésima observação, $i = 1, 2, ..., n$;
- β_0 e β_1 são os parâmetros da equação (coeficientes);
- X_i é o valor da i-ésima observação da variável explicativa; e

· ϵ_i é o erro aleatório na i-ésima observação, com distribuição $N(0, \sigma^2)$.

O parâmetro β_0, também chamado de coeficiente linear, representa o ponto em que a reta de regressão intercepta o eixo y quando $x = 0$. Já β_1 aponta a inclinação da reta, que indica a mudança média da distribuição de y para cada deslocamento em uma unidade da variável explicativa (KUTNER *et al.*, 2005). A Figura 1 apresenta a representação geométrica de uma regressão linear simples.

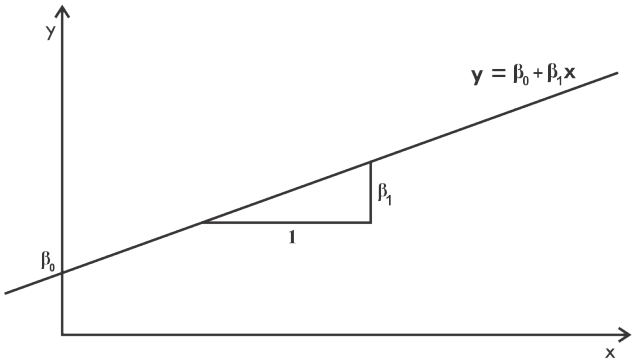

Figura 1: Representação geométrica da reta de regressão linear simples.

2.1.2 Regressão Linear Múltipla

A Regressão Linear Múltipla é baseada na previsão de uma variável resposta com base em mais de uma variável explicativa. Neste caso, o modelo preditivo múltiplo pode ser representado como

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i, \tag{2.1.2}$$

em que:

· Y_i é o valor da variável resposta ou dependente na i-ésima observação, $i = 1, 2, ..., n$;

· $\beta_0, ..., \beta_k$ são os parâmetros da equação (coeficientes);

· $X_{1i}, X_{2i}, ..., X_{ki}$ são os valores da i-ésima observação das k variáveis explicativas; e

· ϵ_i é o erro aleatório na i-ésima observação, com distribuição $N(0, \sigma^2)$.

Assim, em geral o modelo de Regressão Linear pode ser representado em termos da normalidade do erro, o que implica que as observações Y_i sejam variáveis normais independentes, com média $E(Y_i) = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}$ e variância constante σ^2 (KUTNER *et al.*, 2005), isto é,

$$Y_i|(\beta_0, \beta_1, ..., \beta_k, \sigma^2) \sim N(E(Y_i), \sigma^2). \tag{2.1.3}$$

2.1.3 Transformação da Variável Resposta

Quando a variável resposta apresenta diferentes padrões de variação e inconsistência nos pressupostos, como normalidade e heterocedasticidade dos resíduos, é útil transformá-la para atender os pressupostos do modelo. A transformação de Box-Cox é a mais conhecida e pode ser sintetizada pela transformação

$$Y_\lambda = \begin{cases} \log(Y), & \text{se } \lambda = 0 \\ \frac{Y^\lambda - 1}{\lambda}, & \text{se } \lambda \neq 0, \end{cases} \tag{2.1.4}$$

em que o parâmetro de potência λ é sugerido para garantir a adequação dos pressupostos (KUTNER *et al.*, 2005).

2.1.4 Medidas de Comparação entre Modelos

São várias as possíveis medidas de comparação entre modelos estatísticos. Dentre elas, algumas das mais conhecidas são (KUTNER *et al.*, 2005; HARRELL *et al.*, 2015):

Coeficiente de Determinação ($\mathbf{R^2}$): proporção da variabilidade da variável resposta que é explicada pelo modelo, em que quanto mais próximo de 1, melhor é a capacidade do modelo em explicar a variabilidade dos dados. O coeficiente de determinação é definido por

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}. \tag{2.1.5}$$

Erro Quadrático Médio (EQM): média dos quadrados das diferenças entre os valores previstos pelo modelo e os reais (Equação 2.1.6). Neste caso, a qualidade e precisão do modelo é verificada com base nos dados observados e, quanto menor o valor do EQM, melhor é a capacidade de ajuste do modelo. O erro quadrático médio é definido por

$$\mathrm{EQM} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2. \tag{2.1.6}$$

Erro Absoluto Médio (EAM): média das diferenças absolutas entre os valores previstos pelo modelo e os reais (Equação 2.1.7). Quanto menor o valor do EAM, melhor é a capacidade do modelo em prever os valores reais, indicando um ajuste mais preciso aos dados. O erro absoluto médio é definido por

$$\mathrm{EAM} = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|. \tag{2.1.7}$$

Critério de Informação de Akaike (AIC - *Akaike Information Criterion*): medida que considera a complexidade e adequação dos modelos. Quanto menor o valor de AIC, melhor o ajuste do modelo. O AIC é obtido com base na função de verossimilhança (L) e na quantidade de parâmetros (u) estimados do modelo, conforme

$$\text{AIC} = 2u - 2\ln(L). \tag{2.1.8}$$

Nas Equações (2.1.5) a (2.1.8), tem-se que:

- Y_i são os valores observados da variável resposta para cada observação i;
- $\hat{Y}_i$ são os valores previstos do modelo para cada observação i;
- $\bar{Y}$ é a média dos valores observados;
- n é a quantidade de observações;
- u é a quantidade de parâmetros estimados do modelo; e
- L é a função de verossimilhança obtida a partir do modelo ajustado.

2.2 Regressão Logística

A Regressão Logística é um método utilizado para relacionar variáveis respostas de natureza qualitativa nominal ou ordinal, com variáveis explicativas.

2.2.1 Regressão Logística Dicotômica

Também conhecida por Regressão Logística Clássica, Agresti (2007) apresenta a metodologia estatística de Regressão Logística Dicotômica como a previsão de variáveis respostas binárias, isto é, possuem apenas dois valores, comumente orientados por 'sucesso' ou 'fracasso', a partir de variáveis explicativas. Em geral, não há linearidade na relação entre uma variável de resposta binária Y e variáveis explicativas X, então algumas adaptações devem ser consideradas para a adequação do modelo.

A variável Y tem distribuição Bernoulli, com probabilidade de sucesso $\text{P(Y} = 1) = \pi$ e fracasso $\text{P(Y} = 0) = 1 - \pi$. Logo, a média será expressa por $\text{E(Y)} = \pi$ e a variância por $\text{Var(Y)} = \pi(1 - \pi)$. A fim de estimar valores no intervalo (0,1), tem-se definida a função de resposta logística por

$$E(Y_i|x_i) = \pi(x_i) = \frac{\exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki})}. \tag{2.2.1}$$

E assim, utiliza-se uma função de ligação (logito, probit e log-log, por exemplo) para relacionar a média da variável resposta com a parte sistemática do modelo. A função logito é uma das mais usadas neste processo, sendo expressa por

$$\text{logito}[\pi(x_i)] = \log\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki}, \tag{2.2.2}$$

como retrato conveniente da forma aditiva do modelo, que propõe uma representação linear. A interpretação para os coeficientes do modelo é baseada na razão de chances (AGRESTI, 2007), em que:

- $\pi(x_i)$ é a probabilidade de 'sucesso' para a observação i, na presença das variáveis explicativas $x_i = (x_{1i}, x_{2i}, ..., x_{ki})$;
- β_0 é o parâmetro que representa o intercepto da curva logística no eixo y;
- $\beta_1, ..., \beta_k$ são os coeficientes de regressão associados às k variáveis explicativas do modelo; e
- $x_i = (x_{1i}, x_{2i}, ..., x_{ki})$ é o vetor da i-ésima observação das k variáveis explicativas.

A Figura 2 retrata a representação geométrica de uma regressão logística simples.

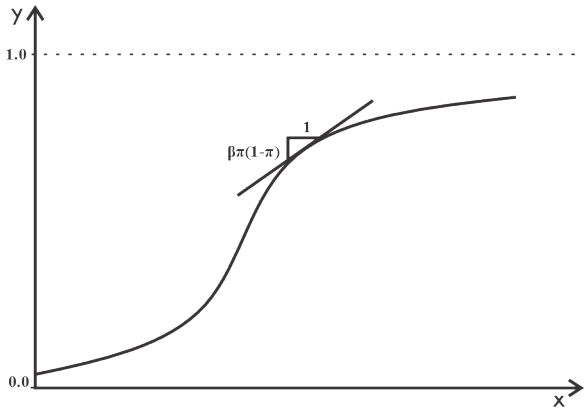

Figura 2: Aproximação linear para a curva de regressão logística, em que $y = \pi(x)$.

2.2.2 Regressão Logística Multicategórica Nominal

Esta regressão tem por finalidade estimar respostas com mais de uma categoria nominal associada, modeladas por uma ou mais variáveis explicativas. A Regressão Logística Multicategórica Nominal é considerada nominal por não haver ordem especificada e relevante entre as categorias da variável resposta (AGRESTI, 2018).

Seja π_j a probabilidade de resposta da categoria j ($j = 1, 2, ..., J$, com $\sum_{j=1}^{J} \pi_j = 1$), e J a quantidade de categorias para Y. Com n observações independentes, o número de

resultados de todas as categorias segue distribuição multinomial. Para J categorias, tem-se $J(J-1)/2$ pares de categorias a serem comparadas, e a mesma quantidade de preditores lineares. O modelo consiste na escolha arbitrária de uma categoria de referência, e todas as demais categorias são comparadas a ela. Desse modo, calcula-se a chance da resposta pertencer a determinada categoria, ao invés de outra (KUTNER *et al.*, 2005; AGRESTI, 2007).

Ao considerar comparações feitas da categoria j ($j = 1, 2, ..., J-1$) com a categoria J, tratando essa última como referência, a função logito nesse caso pode ser expressa como

$$\text{logito}[\pi_{ji}] = \log\left[\frac{\pi_{ji}}{\pi_{Ji}}\right] = \alpha_j + \beta_1 x_{1i} + ... + \beta_k x_{ki}. \tag{2.2.3}$$

Com isso, a probabilidade de resposta da categoria j ($j = 1, 2, ..., J-1$) para a observação i é representada por

$$\pi_{ji} = \frac{\exp(\alpha_j + \beta_1 x_{1i} + ... + \beta_k x_{ki})}{1 + \sum_{j=1}^{J-1} \exp(\alpha_j + \beta_1 x_{1i} + ... + \beta_k x_{ki})}, \tag{2.2.4}$$

e, por consequência, a probabilidade de resposta para a categoria de referência J é obtida conforme

$$\pi_{Ji} = 1 - \sum_{j=1}^{J-1} \pi_{ji}. \tag{2.2.5}$$

Em (2.2.4):

- α_j é o intercepto da curva logística j;
- $\beta = (\beta_1, \beta_2, ..., \beta_k)$ é o vetor de coeficientes de regressão; e
- $x_i = (x_{1i}, x_{2i}, ..., x_{ki})$ é o vetor da i-ésima observação das k variáveis explicativas, $i = 1, 2, ..., n$. Sendo n o número de observações.

Note que os coeficientes de regressão nas Expressões (2.2.3) e (2.2.4) não apresentam índice j, ou seja, este modelo pressupõe que os efeitos das variáveis explicativas são os mesmos em todas as $J-1$ comparações dos pares de categorias. Isso resulta em um modelo mais parcimonioso. Neste caso, os coeficientes da Regressão Logística Multicategórica Nominal são interpretados segundo as razões de chances, com atenção a qual categoria de referência as demais categorias estão sendo comparadas (AGRESTI, 2018).

2.2.3 Regressão Logística Multicategórica Ordinal

A metodologia de Regressão Logística Ordinal considera a ordem das categorias como relevante para a análise. Neste caso, a variável resposta apresenta mais de duas categorias ordenadas, e poderá ser prevista por uma ou mais variáveis explicativas. Variáveis com respostas ordinais podem também ser estudas com as técnicas de regressão logística nominal, no entanto Agresti (2018) orienta:

> *Quando as categorias de resposta são ordenadas, as funções podem utilizar a ordenação. Isso resulta em modelos que têm menos parâmetros, e potencialmente maior poder e interpretações mais simples, quando comparados com modelos logísticos de respostas nominais.*

Três abordagens principais são mencionadas em Agresti (2007) para a elaboração de modelos logísticos ordinais, com base nos logitos: modelos logitos cumulativos; modelos logitos de categorias adjacentes; e modelos logitos de razão sequencial. Qualquer das três abordagens podem ser utilizadas, com a devida mudança na interpretação dos resultados obtidos. Como os modelos cumulativos são mais frequentes em análises estatísticas, e de mais fácil implementação computacional, a revisão principal será condicionada por essa metodologia.

2.2.3.1 Modelos Cumulativos

Considere uma variável resposta Y com J categorias ordenadas. Seja $\pi_j = P(Y = j)$, com $j = 1, 2, ..., J$, então a probabilidade acumulada para Y será a probabilidade dele ser menor ou igual a determinado valor, conforme

$$P(Y \leq j) = \pi_1 + ... + \pi_j, \;\; j = 1, ..., J. \tag{2.2.6}$$

Assim, as probabilidades cumulativas refletem a ordenação das categorias, em que $P(Y \leq 1) \leq P(Y \leq 2) \leq ... \leq P(Y \leq J) = 1$. As funções logitos cumulativas para as $J - 1$ probabilidades iniciais são

$$\text{logito}[P(Y \leq j)] = \log\left[\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right] = \log\left[\frac{\pi_1 + ... + \pi_j}{\pi_{j+1} + ... + \pi_J}\right], \;\; j = 1, ..., J - 1. \tag{2.2.7}$$

Por exemplo, para $J = 3$, tem-se 2 logitos, ambos considerando todas as categorias de respostas:

$$\text{logito}[P(Y \leq 1)] = \log\left[\frac{\pi_1}{\pi_2 + \pi_3}\right], \text{ e}$$

$$\text{logito}[P(Y \leq 2)] = \log\left[\frac{\pi_1 + \pi_2}{\pi_3}\right].$$

Os logitos podem também ser representados por $\text{logito}[P(Y \leq j|x)] = \alpha_j + \beta_1 x_1 + ... + \beta_k x_k$. Isto posto, as probabilidades de resposta podem ser obtidas conforme o modelo de chances proporcionais

$$P(Y_i \leq j|x_i) = \frac{\exp(\alpha_j + \beta_1 x_{1i} + ... + \beta_k x_{ki})}{1 + \exp(\alpha_j + \beta_1 x_{1i} + ... + \beta_k x_{ki})}, \quad j = 1, ..., J-1, \tag{2.2.8}$$

em que α_j representa o intercepto da curva logística do j-ésimo logito, $\beta = (\beta_1, \beta_2, ..., \beta_k)$ o vetor de coeficientes de regressão e $x_i = (x_{1i}, x_{2i}, ..., x_{ki})$ o vetor da i-ésima observação das k variáveis explicativas.

Assim como no modelo (2.2.1), o modelo de regressão logística ordinal (2.2.8) considera fixos os coeficientes de regressão, pressupondo que os efeitos das variáveis explicativas sejam iguais em todos os $J-1$ logitos. Dessa forma, para todo $j \leq (J-1)$, o modelo de chances proporcionais implica que, para cada unidade acrescida a variável explicativa x_s $(s = 1, 2, ..., k)$, a chance (em relação a categoria de referência J) da observação pertencer à j-ésima categoria ou uma categoria inferior é um múltiplo de e^{β_s}. Ademais, a função ordinal (2.2.8) permite estimar o logaritmo da probabilidade de a variável resposta atribuir os valores de classes inferiores ou iguais a j, comparativamente com a probabilidade de tomar os valores das classes superiores a j (CELLA, nov. 2013; HARRELL *et al.*, 2015; AGRESTI, 2018).

A título de exemplo, a Figura 3 representa de forma gráfica, respectivamente, as probabilidades para cada categoria de uma variável resposta e as probabilidades cumulativas do modelo logito, para $J = 4$.

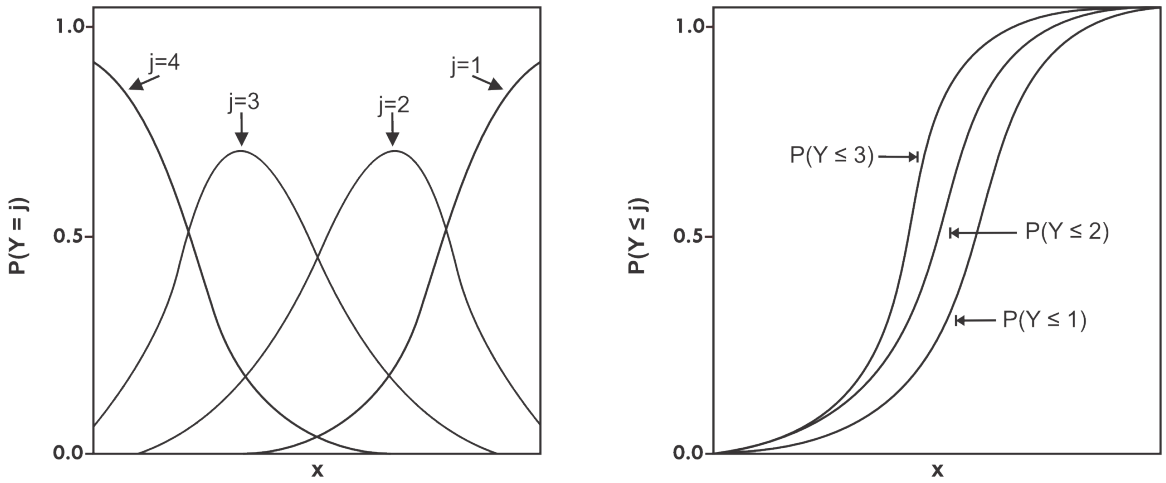

Figura 3: Representações respectivas de: probabilidades para cada uma das 4 categorias no modelo logito cumulativo; e probabilidades cumulativas no modelo logito cumulativo para variável resposta com 4 categorias.

2.3 Regressão de Poisson

A Regressão de Poisson possibilita analisar a previsão de uma variável resposta com caráter de contagem (valores discretos e positivos) ou taxa, a partir de uma ou mais variáveis explicativas (AGRESTI, 2007). Trata-se de uma das famílias dos Modelos Lineares Generalizados (MLG), que representam o conjunto de modelos lineares com uma distribuição da família exponencial (SCHMIDT, 2003).

A função mais comum utilizada na modelagem de Poisson contempla o logaritmo da média. Nesse caso, o modelo log-linear de Poisson assume distribuição Poisson para Y, e se baseia na função de ligação logarítmica. Para uma única variável explicativa, o modelo de Poisson log-linear pode ser expresso como (AGRESTI, 2018)

$$\log(\mu) = \alpha + \beta x, \tag{2.3.1}$$

em que μ é a média da distribuição que satisfaz a relação exponencial

$$\mu = \exp(\alpha + \beta x). \tag{2.3.2}$$

Agresti (2018) esclarece que o aumento em uma unidade em x tem o impacto multiplicativo de e^{β} em μ. Se $\beta > 0$, então $e^{\beta} > 1$, e a média de Y aumenta a medida que x aumenta. O contrário ocorre se $\beta < 0$, nesse caso a média de Y diminui a medida que x aumenta. Caso β seja nulo, a média de Y não altera com a mudança em x.

Ao estender para um modelo com mais de uma variável explicativa, a função de ligação logarítmica pode ser escrita como na Equação (2.3.3) e a média μ como em (2.3.4)

(KUTNER *et al.*, 2005):

$$\log(\mu_i) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} \tag{2.3.3}$$

$$\Downarrow$$

$$\mu_i = \exp(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki}), \tag{2.3.4}$$

de modo que α representa o intercepto da curva log-linear, $\beta = (\beta_1, \beta_2, ..., \beta_k)$ o vetor de coeficientes de regressão e $x_i = (x_{1i}, x_{2i}, ..., x_{ki})$ o vetor da i-ésima observação das k variáveis explicativas.

No modelo de Poisson, é pressuposto que a variável resposta segue distribuição Poisson, ou seja, a média da variável resposta deve ser igual à sua variância. Porém, é comum que para dados experimentais essa condição seja violada, gerando superdispersão, quando a variância é maior que a média, ou subdispersão, quando a variância é menor que a média. Dessa forma, a aplicação da Regressão de Poisson será possível ao considerar alguns ajustes (CONSUL; FAMOYE, 1992; SCHMIDT, 2003).

A Figura 4 representa a distribuição da variável resposta Y para diferentes valores de μ. Observe que, ao passo que o valor de μ aumenta, a curva se achata, desloca-se para a direita e a distribuição discreta de Poisson se aproxima cada vez mais de uma distribuição normal (KUTNER *et al.*, 2005).

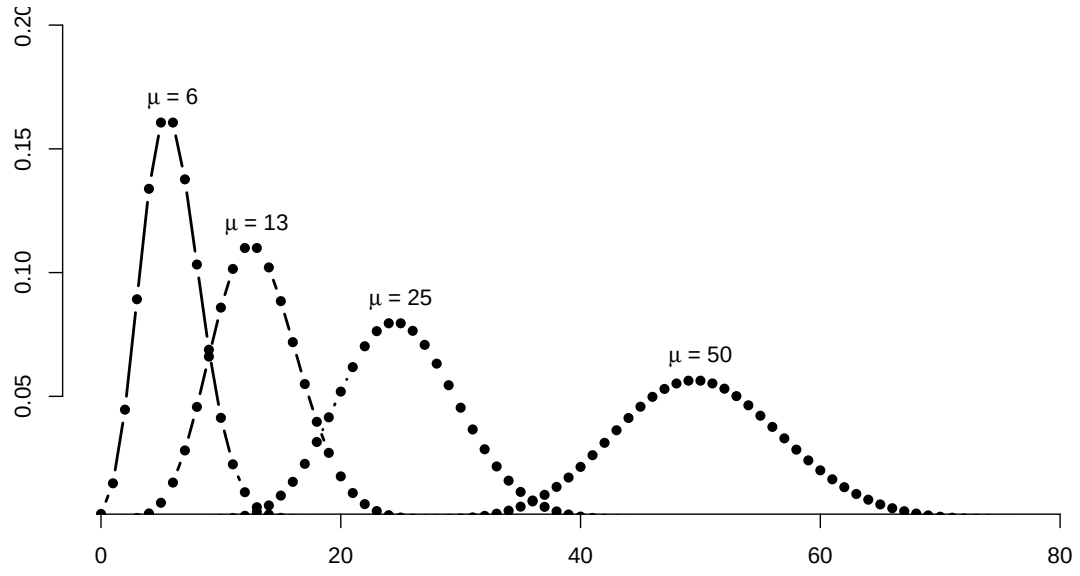

Figura 4: Distribuição de Poisson para diferentes médias μ.

3 Técnicas de Imputação

As análises estatísticas conhecidas são formuladas para estudar bancos de dados completos. Com isso, o tratamento adequado de situações onde existem dados faltantes pode ser o diferencial ao concluir sobre uma investigação (HARRELL *et al.*, 2015).

3.1 Dados Faltantes

Os dados faltantes ou *missings* são problemas recorrentes em processos de pesquisa, e apenas desconsiderá-los ou utilizar técnicas simples, como imputar por medidas de tendência central, pode não ser a melhor opção. Por exemplo, Harrell *et al.* (2015) explicam que se uma variável apresenta dados faltantes e a respectiva observação é excluída do ajuste de um modelo estatístico, tal feito possibilita gerar coeficientes extremamente viesados e/ou imprecisos.

Vários motivos podem gerar não-resposta, e por consequência perda de informação em um conjunto de dados. A classificação sobre o que gera dados faltantes foi explorada por vários autores. Schafer e Graham (2002) identificam a causa da ausência de dados por padrões de não-resposta. Já McKnight *et al.* (2007) justificam conforme a fonte que gerou o problema, em geral explicada por razões de um participante envolvido em pesquisa. Rubin e Wiley (1987) propõem uma abordagem mais completa e aceita por cientistas sobre essa questão, e descreve que o motivo da falta de dados pode ser resumido por mecanismos.

Segundo Rubin e Wiley (1987), são três os mecanismos que explicam a ausência de dados: dados faltantes completamente ao acaso (MCAR, *Missing Completely at Random*), dados faltantes ao acaso (MAR, *Missing at Random*) e dados faltantes não aleatórios (MNAR, *Missing not at Random*).

Faltantes completamente ao acaso (MCAR): o motivo não está relacionado às variáveis observadas no estudo, sejam elas as que apresentam os valores faltantes ou não. Neste caso, por exemplo, em uma pesquisa de opinião a ausência de dados não está relacionada às respostas dos entrevistados. Assim, considera-se que a probabilidade do dado faltante ocorrer é a mesma para todas as observações.

Faltantes ao acaso (MAR): ocorrem quando o padrão de não-resposta pode ser explicado apenas por outras variáveis do banco dados, logo o motivo da ausência não é motivado pela variável que apresenta dados faltantes.

Faltantes não aleatórios (MNAR): ocorrem quando a ausência está diretamente relacionada aos valores da própria variável analisada. Dessa forma, o motivo para

dados incompletos está relacionado a valores não observados. Diferente dos mecanismos MCAR e MAR, os faltantes não aleatórios não podem ser completamente ignorados em uma análise.

3.2 Imputação

O tipo de imputação de dados faltantes pode ser orientado de diversas formas, porém é possível dividir tais técnicas em dois grupos: imputação única ou simples, ao imputar os valores uma única vez; e imputação múltipla, ao imputar com base em um processo iterativo. Ambos os métodos se complementam, uma vez que fundamentos utilizados na imputação única são considerados no processo múltiplo, e, a depender do cenário e investigação da pesquisa, os dois tipos de imputação podem ser ponderados em uma mesma análise. (RUBIN, 1996; NUNES; KLÜCK; FACHEL, 2010).

3.2.1 Imputação Única

A imputação única ocorre quando a recuperação dos dados faltantes sucede por apenas uma única vez, ao utilizar principalmente técnicas como substituição por: medidas de tendência central (média, mediana e moda); valores de outras observações semelhantes; estimativas de máxima verossimilhança; último ou próximo valor observado; e respostas únicas obtidas por modelos de regressão.

No contexto de modelagem estatística, com a imputação única obtida por modelo de regressão, a variável que apresenta dados faltantes é considerada a variável resposta do modelo, e assim seus valores são imputados uma única vez com base nas demais variáveis completas presentes no mesmo conjunto de dados. Neste caso, as covariáveis presentes no modelo ajustado são prioritariamente significativas e todos os pressupostos devem ser verificados na análise de diagnóstico final, para a melhor imputação dos valores.

Apesar de certas limitações referentes à imputação única, estudos apontam que em muitos casos o desenvolvimento de técnicas mais simples podem trazer grande impacto no estudo em questão, sendo muito mais eficiente do que apenas desconsiderar as variáveis que apresentam dados faltantes (ENGELS; DIEHR, 2003; HARRELL *et al.*, 2015; NUNES, 2007).

3.2.2 Imputação Múltipla

A imputação múltipla ocorre conforme um processo iterativo. Assim, consiste na obtenção de M bancos de dados completos, de modo que a substituição dos valores faltan-

tes seja realizada M vezes. Todos os bancos gerados são assim analisados, a fim de gerar estimativas pontuais como média e erro padrão. Ao presumir que os dados ausentes são gerados ao acaso, os resultados das estimativas alcançadas não serão viesados. A complexidade por trás do método de imputação múltipla exige, em muitas situações, a utilização de técnicas mais sofisticadas, como: modelagem estatística (análise de regressão), Média Preditiva e Cadeias de Markov via Monte Carlo (MCMC) (RUBIN, 1996; SCHAFER; GRAHAM, 2002; NUNES, 2007; VINHA, 2016). Apesar disso, Buuren (2018) menciona que a imputação múltipla é atualmente considerada o método mais eficaz para lidar com dados incompletos em muitas áreas de pesquisa.

No contexto de modelagem estatística, Harrell *et al.* (2015) explicam:

> *Quando um modelo de regressão é usado para imputação, o processo envolve adicionar um resíduo aleatório ao 'melhor palpite' para os valores ausentes, a fim de produzir a mesma variância condicional que a variável original.* [...] *Cada repetição resulta em um conjunto de dados 'completado' que é analisado usando o método padrão. As estimativas dos parâmetros são médias dessas múltiplas imputações para obter estimativas melhores do que aquelas de imputação única.*

A quantidade de imputações M é comumente decidida com base na proporção de dados faltantes da variável incompleta. White, Royston e Wood (2011) sugerem o uso de $M = 100p$ imputações, sendo p a proporção de dados faltante. Buuren (2018) e Royston (2004) recomendam um valor mínimo de 20 imputações, e mencionam que é importante considerar um limite eficiente de imputações, para não sobrecarregar, de maneira desnecessária, o 'custo' computacional para gerar os resultados. Já Bodner (2008) propõe que, para proporções de dados faltantes p iguais a 0.05, 0.1, 0.2, 0.3, 0.5, 0.7 e 0.9, são recomendados valores de M iguais a 3, 6, 12, 24, 59, 114 e 258, respectivamente. Para valores intermediários, uma interpolação pode ser adotada. De toda forma, a decisão sobre a quantidade de imputações escolhida deve considerar também a complexidade do conjunto de dados e o motivo que gerou as variáveis incompletas.

3.3 Técnicas

A literatura aponta grande quantidade de possíveis técnicas para o tratamento de valores ausentes. Muitos analistas, por falta de proximidade com as técnicas, baseiam-se em métodos como substituição por média/moda ou eliminação das observações. No entanto, tal simplicidade na maioria dos casos não se mostra a mais adequada, e o uso de de técnicas mais sofisticados como regressão estatística pode ser o diferencial no processo de imputação de dados faltantes (MCKNIGHT *et al.*, 2007; NUNES, 2007; VINHA, 2016).

3.3.1 Medidas de Tendência Central

As informações ausentes são substituídas com base nas estatística descritivas média, moda ou mediana. A imputação ocorre conforme a natureza da variável que apresenta a falta de dados. Além disso, há possibilidade de imputar os resultados dos valores observados da variável incompleta ou utilizar estatísticas de tendência central de outra variável com resultados similares (ENDERS, 2010; NUNES, 2007; VINHA, 2016). Assim, a imputação pela:

- **Média**: é utilizada em variáveis de natureza quantitativa discreta ou contínua. Os dados faltantes são imputados pelo valor médio dos dados restantes da variável incompleta ou da variável com característica similar;
- **Mediana**: é utilizada em variáveis de natureza quantitativa ou qualitativa ordinal. Para característica quantitativa, a mediana será mais eficiente que a média caso tenha valores extremos na variável de interesse. Os dados faltantes são imputados pelo valor da mediana dos dados restantes da variável incompleta ou da variável com característica similar;
- **Moda**: é utilizada em variáveis com natureza qualitativa nominal ou ordinal, em que a categoria com maior frequência é escolhida. Os dados faltantes são imputados pela moda, ou seja, valor ou categoria de maior frequência presente nos dados restantes da variável incompleta ou da variável com característica similar.

3.3.2 Modelagem Preditiva

As técnicas de imputação com uso de modelagem preditiva apresentam robustez (resultados confiáveis e precisos) e consistência (resultados com qualidade e eficiência em diferentes situações), quando comparadas com substituição por medidas de tendência central. À priori, a natureza e características da variável a ser imputada define qual regressão estatística será indicada ao estudo. No contexto de imputação, as regressões lineares, logísticas e de Poisson são suficientes para analisar a maioria dos bancos de dados, pois abrangem todas as possíveis classes da variável resposta Y, que pode assumir perfil quantitativo (discreto ou contínuo) ou qualitativo (nominal ou ordinal) (BUCK, 1960; KUTNER *et al.*, 2005; AMBLER; OMAR; ROYSTON, 2007).

Em sua tese, Vinha (2016) acrescenta alguns pontos relacionados à imputação com uso de regressão, baseados em Peugh *et al.* (2004) e Enders (2010):

> Quando muitas variáveis apresentam dados ausentes esse método perde a atratividade, pois o número de equações a serem estimadas é grande (ENDERS, 2010). Como a substituição é feita por um valor predito baseado nas relações entre as variáveis, o valor substituído está exatamente na linha que descreve a relação, o que pode resultar na superestimação das covariâncias entre as variáveis (PEUGH; ENDERS, 2004). Por outro lado, a subestimação da variabilidade é menos acentuada do que a observada com a utilização da imputação pela média e o procedimento gera estimativas não viesadas quando os dados são do tipo MAR (Missing at Random).

Os dados ausentes são imputados com uso dos valores previstos por cada regressão estatística especificada. Assim, a escolha de cada técnica de imputação fundamentada por modelagem preditiva pode ser simplificada por:

- **Regressão Linear**: para variáveis respostas com natureza quantitativa contínua ou discreta (casos específicos);
- **Regressão Logística Dicotômica**: para variáveis respostas com natureza qualitativa binária ('sucesso' ou 'fracasso'), em geral caracterizadas por valores 0 e 1;
- **Regressão Logística Multi-categórica Nominal**: para variáveis respostas com natureza qualitativa, que apresentam mais do que duas categorias não ordenadas;
- **Regressão Logística Multi-categórica Ordinal**: para variáveis respostas com natureza qualitativa, que apresentam mais de duas categorias associadas, com ordem especificada e relevante ao estudo;
- **Regressão de Poisson**: para variáveis respostas com natureza quantitativa discreta, em geral com caráter de contagem.

Devido a questões relacionadas ao conhecimento de métodos estatísticos e à área de imputação de dados da maioria dos pesquisadores, a escolha da técnica conforme a natureza da variável resposta pode variar sem grande perda nos resultados obtidos. Tal fato se aplica, por exemplo, ao contexto em que a variável de interesse apresenta característica de contagem. Neste caso, apesar de ser preferível e recomendável ajustar um modelo de regressão de Poisson para o processo de imputação, talvez seja suficiente imputar os dados faltantes com ajuste de um modelo linear em um processo de imputação múltiplo, devido à maior familiaridade com análise de regressão linear (HARRELL *et al.*, 2015; BUUREN, 2018).

Uma das desvantagens ao utilizar modelagem preditiva é que indícios apontam que imputação via regressão aumentam artificialmente as relações nos dados, levando as correlações a serem tendenciosas para cima e a variabilidade a ser subestimada. As imputações neste casos são excessivamente precisas, o que pode resultar em relações falsas

positivas e espúrias, fatores que devem ser considerados em processos de estudos (BUUREN, 2018).

3.3.3 Outras Alternativas

Algumas outras técnicas que envolvem imputação única e múltipla também são conhecidas e validadas no meio científico.

Quando há dados faltantes em uma pesquisa, é possível selecionar valores de respondentes que sejam similares em relação a variáveis auxiliares para a imputação. Neste caso, a técnica conhecida por *hot deck* pode ser bem implementada. Os respondentes são chamados de 'doadores' e, para selecioná-los, é preciso localizar o indivíduo com dado observado mais parecido com o indivíduo com dado faltante em relação às variáveis auxiliares. E assim substituir o dado faltante pelo valor do respondente correspondente (RUBIN, 1996).

Em situações que há mais de um respondente pareado em uma pesquisa com dados faltantes, é possível utilizar o método de imputação que considera a observação vizinha mais próxima. Comumente utilizado em pesquisas da área da saúde, esse método cria um critério de classificação para identificar o registro mais parecido com aquele que possui o dado faltante, e assim esse registro se torna o valor que será imputado (NUNES, 2007).

Em bases incompletas com dados longitudinais, ou seja, valores ao longo do tempo de uma mesma unidade ou indivíduo, algumas técnicas devem ser consideradas para solucionar o problema de dados faltantes. Por se tratar de série temporal, além da imputação por medidas de tendência central com análise prévia ou posterior da observação ausente, é possível utilizar valores da próxima observação da unidade/indivíduo para a substituição. Nessa perspectiva, há também a substituição de cada dado faltante pela média do valor anterior e posterior da observação ausente, ou até mesmo é válida a substituição do dado faltante pelo último valor observado. Diversos métodos para dados longitudinais são sugeridos por diferentes autores, dependendo da área de estudo. É importante discutir e considerar esses métodos de acordo com os resultados desejados (ENGELS; DIEHR, 2003; VINHA, 2016).

3.3.4 Avaliação de Desempenho

A qualidade da técnica de imputação deve ser avaliada para obter inferências estatisticamente válidas a partir de dados faltantes. São várias as medidas capazes de informar sobre a validade estatística de cada técnica implementada, cabendo ao pesquisador decidir qual a melhor métrica para comparar os métodos de imputação sugeridos.

Contudo, Buuren (2018) adverte que:

> *Avaliar a discrepância entre os dados verdadeiros e os dados imputados pode parecer uma maneira simples e atraente de selecionar o melhor método de imputação. No entanto, não é útil avaliar métodos com base apenas em sua capacidade de recriar os dados verdadeiros. Pelo contrário, selecionar tais métodos pode ser prejudicial, pois eles podem aumentar a taxa de falsos positivos. Imputação não é previsão.*

Em vista disso, é possível avaliar o desempenho de técnicas de imputação sem comparar de forma direta os dados verdadeiros com os imputados, mas sim comparar a influência de tais imputações em determinados cenários.

Em um processo de simulação controlado a decisão sobre o método de comparação se torna mais fácil e eficiente. Habitualmente, existem dois mecanismos que afetam os dados observados: o mecanismo de amostragem e o mecanismo de dados faltantes. A simulação pode abordar separadamente o mecanismo de amostragem, o mecanismo de dados faltantes e ambos os mecanismos combinados. Exemplos de estimativas científicas conhecidas usadas para comparar técnicas de imputação incluem média, covariância, correlações e coeficientes de regressão referentes ao conjunto de dados completo (BUUREN, 2018).

Por exemplo, em uma simulação para imputações aplicadas em variáveis quantitativas, é plausível usar um modelo preditivo linear como base comparativa e de referência, com medidas de ajuste obtidas a partir dos dados completos (como Critério de Informação de Akaike - AIC e Erro Quadrático Médio - EQM). Neste caso, novas medidas podem ser geradas após os valores serem imputados, e por fim comparadas com os valores reais de AIC e EQM. Sendo assim possível avaliar o desempenho de diferente técnicas em um mesmo cenário, sem comparar diretamente os valores reais com os imputados, mas sim com auxílio de estimativas científicas como referência.

4 Aplicações

A imputação de dados propõe diversas possibilidades, mas é preciso considerar fatores importantes, como: porcentagem de dados faltantes; motivos da ausência da informação; recursos disponíveis; tempo para análise dos modelos propostos; e decisão sobre o tipo de imputação mais adequado para a complexidade do problema. Neste contexto, para ilustrar a teoria por trás deste trabalho, a metodologia foi dividida em três partes principais. Essas partes foram subdivididas em situações que envolvem tanto imputações únicas, quanto imputações múltiplas. O foco maior está em representar possíveis soluções ou algoritmos para cenários recorrentes de dados faltantes, com demonstração em um único conjunto de dados simulados, apresentando métodos de simples a complexos. Toda a análise foi fomentada com uso do software R versão 4.3.2 (R Core Team, 2023).

A primeira parte se baseia inteiramente em técnicas de imputação única simples. Apesar da literatura já apontar bons argumentos a favor dos métodos de modelagem estatística, o objetivo principal está em introduzir a importância do uso de modelagem no contexto de imputação de dados faltantes, apontando resultados que apontam o uso de técnicas mais sofisticadas em contextos reais, a partir de resultados prévios simulados. Inicialmente, o foco está em comparar técnicas de imputação com uso de estatísticas de tendência central e uso de modelagem preditiva, ao retratar suas principais diferenças.

A segunda parte está relacionada à comparação direta da imputação única com uso apenas de modelagem preditiva em um cenário mais complexo, em que mais de uma variável apresenta dados faltantes. Apesar de simular várias bases de dados para fins ilustrativos, a recuperação dos dados faltantes ocorre apenas uma única vez nas partes 1 e 2.

Já a terceira parte se baseia em técnicas de imputação múltipla. Conforme resultados prévios da primeira e segunda parte, que apontam a eficiência do uso de regressão no processo de recuperação de dados, a parte 3 decorre apenas por técnicas com uso de modelagem preditiva. Nesta etapa os valores a serem imputados são gerados mais de uma vez para cada observação e, utilizando-se de métodos estatísticos, tais valores são por fim selecionados. O destaque está nas possibilidades geradas pelo processo iterativo da recuperação de dados faltantes. Ao traçar diferentes cenários, variando escolhas de variáveis e ordens das etapas dos processos, a quantidade de alternativas para a imputação dos dados aumenta de modo considerável.

4.1 Materiais e Métodos

A metodologia e os materiais utilizados nesta pesquisa foram aplicados em todas as partes mencionadas, de modo que elas apresentem um mesmo delineamento para fins comparativos. Portanto, em todos os cenários abordados, a pesquisa foi direcionada frente: ao mesmo banco de dados original e completo; às mesmas bases com dados faltantes simulados; e ao mesmo modelo de regressão linear final para comparação das técnicas de imputação, ou seja, preservando as mesmas covariáveis independentes da técnica utilizada.

4.1.1 Banco de Dados

Composto por 200 observações, o banco de dados utilizado neste trabalho foi simulado frente à geração de números pseudo-aleatórios com base na distribuição desejada atribuída a cada variável. Com 6 variáveis ao todo, caracterizadas por y, x_1, x_2, x_3, x_4 e x_5, cada uma apresenta certa distribuição pré definida, a fim de estudar técnicas de imputação aplicadas a diferentes cenários e naturezas de variáveis. Assim, após fixar uma semente para simulação dos resultados, cada variável pode ser caracterizada da seguinte forma:

- **y** é uma variável quantitativa contínua, com distribuição normal. Representa a variável resposta que será utilizada como critério comparativo das técnicas de imputação apresentadas. Frente ao desvio padrão de 3.5, o ajuste do preditor linear $\mu(Y_i)$ para a i-ésima observação (Equação (4.1.1)) considerou como significativas para o modelo de predição de y apenas as variáveis x_1, x_2, x_4 e x_5. Assim, $y_i \sim N(\mu = \mu(Y_i), \sigma^2 = 3.5^2)$, com

$$\mu(Y_i) = 3 + 7X_{1i} + 8X_{2i} + 0.01X_{3i} + 4X_{4i} + 3X_{5i}. \tag{4.1.1}$$

Em (4.1.1), a variável:

- $\mathbf{x_1}$ é qualitativa dicotômica, com valores 0 ('fracasso') e 1 ('sucesso'), e distribuição de Bernoulli, com probabilidade de sucesso igual a 0.6;
- $\mathbf{x_2}$ é quantitativa contínua, com distribuição normal de parâmetro média igual a 3 e desvio padrão igual a 3;
- $\mathbf{x_3}$ é quantitativa discreta, com característica de contagem e distribuição de Poisson com média 3;
- $\mathbf{x_4}$ é quantitativa contínua, com distribuição normal com desvio padrão igual a 2 e parâmetro média gerado com auxílio de preditor linear $\mu(X_{4i})$ considerando as

variáveis x_1 e x_2, para cada i-ésima observação. De modo que

$$\mu(X_{4i}) = 1 + 2X_{1i} - 1.5X_{2i}, \tag{4.1.2}$$

induzindo assim certa correlação entre as variáveis para fins de simulação; e

- $\mathbf{x_5}$ é quantitativa discreta, com característica de contagem e distribuição de Poisson com parâmetro média gerado com auxílio de preditor linear $\mu(X_{5i})$ considerando as variáveis x_1 e x_2, para cada i-ésima observação. De modo que

$$\mu(X_{5i}) = 2 + 0.5X_{1i} - 0.5X_{2i}, \tag{4.1.3}$$

induzindo assim certa correlação entre as variáveis para fins de simulação.

4.1.2 Modelo Preditivo Final

Toda a comparação entre as técnicas e os comportamentos dos dados após valores imputados ocorreu conforme adequação de um único modelo preditivo para a variável y. Dessa forma, foi ajustado um modelo de regressão linear considerando y como variável resposta, e todas as demais como covariáveis (Equação (2.1.2)).

Conforme análise prévia, apesar de alguns valores atípicos serem identificados no conjunto de dados, a variável resposta manteve seu aspecto original, sem transformação. Isso se deu devido a critérios de comparação pré selecionados, em que, pela grande quantidade de bases de dados simuladas, o 'custo' computacional para realizar a transformação de Box-Cox precisou ser reavaliado no processo de simulação. Assim, aspectos relacionados a presença de *outliers*, alavancagens e pontos influentes foram considerados, porém tratados como casos atípicos passíveis de impactar qualquer uma das bases obtidas, e portanto não interferiram nas conclusões sobre os desempenhos das técnicas de imputação apresentadas. O estudo referente à identificação de observações de grande influência no modelo final pode ser verificado na Figura 6.

A normalidade dos resíduos do modelo proposto não foi violada, com base no teste de Kolmogorov-Smirnov, que não rejeitou a hipótese de que o resíduo apresenta distribuição normal (p-valor = 0.6270), e também conforme os gráficos da Figura 5 referentes à distribuição e aos quantis normais dos resíduos. Além disso, foi verificada a homoscedasticidade dos resíduos, que foi atestada por meio do teste de Breusch-Pagan que, ao p-valor 0.1965, não apresentou evidência suficiente para rejeitar a hipótese de que a variância dos erros do modelo é constante.

Quanto à multicolinearidade, ao calcular o fator de inflação da variância (VIF - *Variance Inflation Factor*) para cada variável, que mede quanto a variância do coefi-

ciente estimado para uma variável é inflada devido à multicolinearidade com as outras variáveis explicativas, foi verificado valores baixos para todas as covariáveis, indicando não multicolinearidade. No entanto, é válido ressaltar que certa dependência foi induzida previamente entre algumas variáveis explicativas do modelo, mas sem comprometer tanto o ajuste final. Tal fato também foi verificado por gráficos de correlação e coeficientes de Pearson, em que o problema de dependência entre as variáveis não parece ter tanto peso, dada a correlação moderada e não extrema.

Por fim, a baixa exogeneidade do modelo proposto foi atestada com apoio de resultados do teste de Durbin-Watson que, ao p-valor 0.4962 e estatística do teste 1.9989, indica ausência de autocorrelação, sendo mais provável que os erros de medidas não interfiram nas variáveis explicativas. Portanto, com base nas análises dos pressupostos e diagnósticos, a adequação modelo de regressão final foi considerada consistente. A Figura 5 reflete algumas das análises de diagnósticos realizadas.

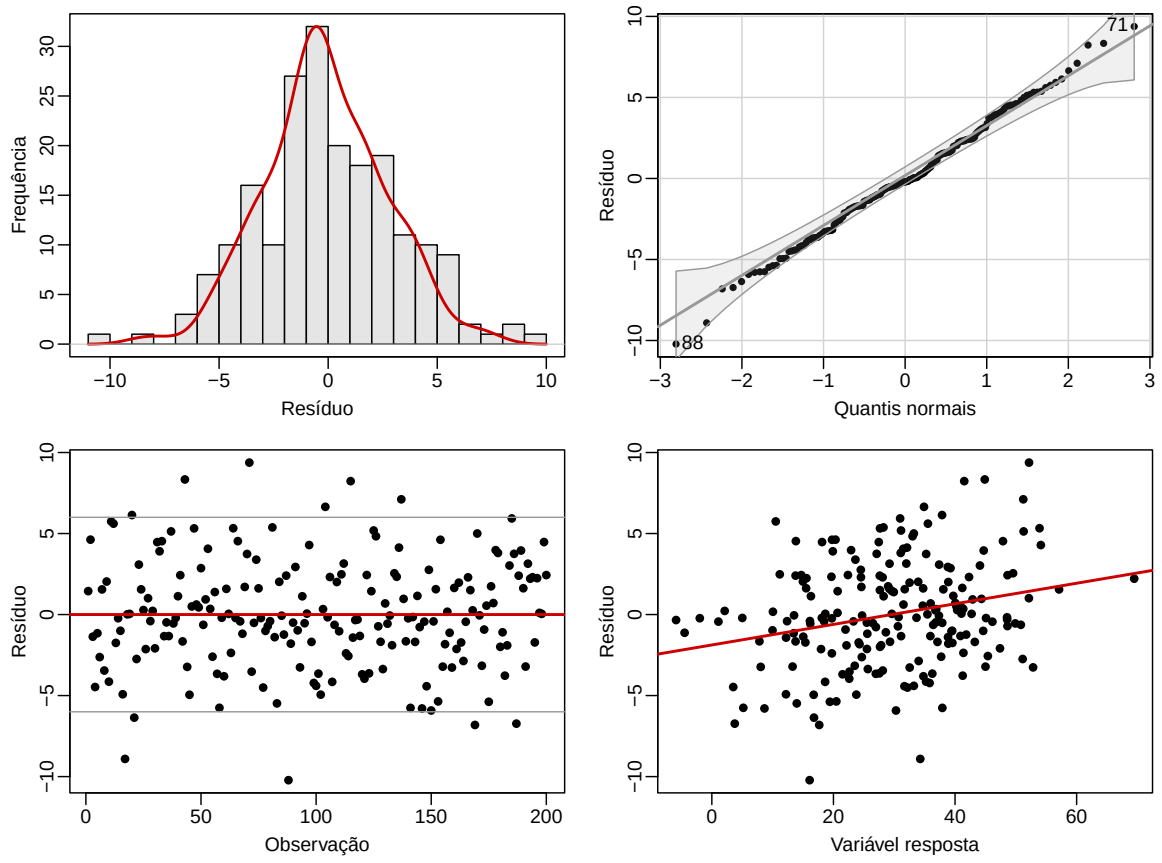

Figura 5: Gráficos de diagnóstico dos resíduos do modelo de regressão linear, para estudo da variável resposta y e das técnicas de imputação. Com representação da distribuição, dos quantis normais, da variabilidade e da correlação com a variável resposta.

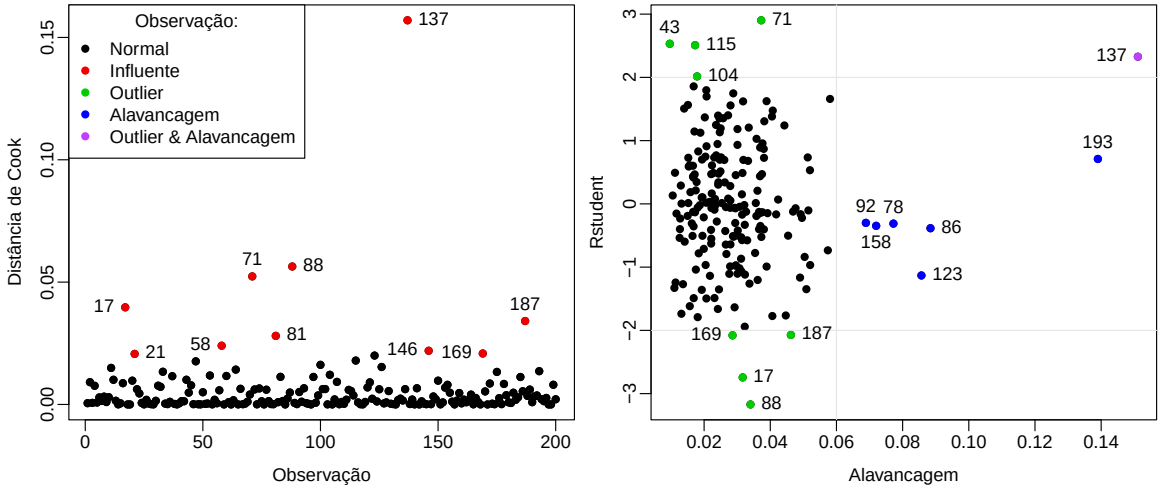

Figura 6: Distância de Cook e gráfico de dispersão (resíduo studentizado vs alavancagem), respectivamente, para análise de observações de grande influência no modelo final para y.

Algumas medidas e coeficientes observados também foram obtidos para fins comparativos, referentes ao modelo linear ajustado aos dados completos antes das simulações, sendo eles: Erro Quadrático Médio (EQM = 11.256), Erro Absoluto Médio (EAM = 2.574) e Coeficiente de Determinação ($R^2 = 0.936$).

4.1.3 Simulação de Dados Faltantes

Devido à completude do banco de dados, antes da aplicação das técnicas, algumas simulações foram feitas para gerar bases com diferentes proporções de dados faltantes em variáveis específicas, e assim comparar resultados de cada método em diferentes cenários. Em todas as etapas o processo de seleção de cada observação com dado faltante ocorreu de forma aleatória, sendo repetido 300 vezes para cada proporção. A análise considerou cinco diferentes proporções de dados faltantes (1%, 5%, 10%, 20% e 40%). Resultando em um total de 1500 bases para o desenvolvimento de cada técnica de imputação, seja ela única ou múltipla.

4.1.4 Critérios de Comparação

Todas as técnicas de imputação apresentadas foram comparadas conforme algumas medidas obtidas pelos modelos preditivos, por sua vez ajustados após o processo de imputação das bases simuladas com dados faltantes. Assim, os parâmetros principais utilizados para comparar as adequabilidades dos modelos lineares finais foram o Erro Quadrático Médio (EQM) de previsão e o Coeficiente de Determinação (R^2).

Neste caso, o EQM e R^2 do modelo ideal, ajustados para o conjunto de dados sem alteração e completo, foram utilizados como referência comparativa. Após cada método realizado e valores dos parâmetros encontrados para cada conjunto de dados simulados, algumas medidas descritivas como média, variância e quantis (2.5%, mediana e 97.5%) foram obtidas para os EQMs e R^2, conforme respectiva proporção de dados faltantes. Espera-se que quanto menor os valores de tendência central e a variabilidade dos EQMs, e quanto maior os resultados para os Coeficientes de Determinação, melhor será a técnica de imputação para ajustar os modelos preditivos para y. Ademais, as análises referentes às métricas citadas também foram complementas com base em *boxplots*, curvas de densidade e testes de hipóteses para comparar amostras.

4.2 Parte 1: Imputação Única em Cenários Simples

O objetivo principal está em verificar a eficiência de três técnicas de imputação única em uma situação mais simples, em que apenas uma variável apresenta dados faltantes, e destacar suas diferenças conforme os critérios mencionados na seção 4.1.4. Assim, apenas a variável x_5 foi selecionada para o processo de simulação, posto que é significativa para o modelo preditivo final. Por ser tratar de valores com natureza quantitativa discreta, as técnicas utilizadas foram imputação por: estatística média; valores previstos por regressão de Poisson; e valores previstos por regressão linear. É importante destacar que as técnicas de imputação por modelagem preditiva só serão mais eficazes do que a imputação por medidas de tendência central se as variáveis utilizadas para os modelos de imputação forem significativas. Caso contrário, a imputação por meio de regressão estatística produzirá resultados semelhantes às técnicas mais simples. Assim a seleção cuidadosa das variáveis explicativas é fundamental para garantir a eficácia da imputação.

4.2.1 Método

Para imputar pela estatística média, o valor médio dos dados não faltantes de x_5 foi obtido e imputado em todas as células que apresentaram dados faltantes.

No caso das técnicas via modelagem preditiva, conforme análise de medidas de comparação entre modelos, tanto para o caso considerando x_5 com distribuição Poisson (modelo de Poisson), quanto considerando com distribuição normal (modelo linear), as variáveis x_1 e x_2 foram selecionadas para imputar os valores ausentes em x_5. Junto a análises de diagnósticos, o Critério de Informação de Akaike (AIC) foi a medida principal usada para selecionar as covariáveis de ambos os modelos.

A Equação (4.2.1) expressa o modelo linear de obtenção dos valores previstos $X_{5,ij}$ para a i-ésima observação com dado faltante em x_5 no j-ésimo banco de dados simulado,

que serão imputados conforme as 1500 regressões lineares:

$$X_{5,ij} = \beta_{0j} + \beta_{1j}X_{1i} + \beta_{2j}X_{2i}. \tag{4.2.1}$$

Já a Equação (4.2.2) se refere ao modelo de Poisson para obtenção dos valores previstos $X_{5,ij}$ para a i-ésima observação com dado faltante em x_5 no j-ésimo banco de dados simulado, que serão imputados conforme as 1500 regressões de Poisson:

$$X_{5,ij} = \exp(\alpha_j + \beta_{1j}X_{1i} + \beta_{2j}X_{2i}). \tag{4.2.2}$$

Em ambos os modelos propostos, X_{1i} e X_{2i} representam as observações da base original (i = 1,2,...,200) para as variáveis x_1 e x_2, respectivamente. Além disso, β_{1j} e β_{2j} representam os coeficientes de regressão dos modelos. Na Equação (4.2.1), β_{0j} representa o intercepto da reta de regressão linear. E na Equação (4.2.2), α_j se refere ao intercepto da curva log-linear da regressão de Poisson. Importante que, apesar de alguns coeficientes retratarem a mesma definição entre os modelos mencionados, os parâmetros e coeficientes das Equações (4.2.1) e (4.2.2) foram estudados separadamente.

Ao simular x_5 como variável resposta com dados faltantes, e x_1 e x_2 como covariáveis completas, tem-se um cenário em que dados de apenas uma variável passou pelo processo de imputação. Com isso os modelos de previsões finais foram ajustados aos novos dados encontrados após o processo de imputação, e EQMs e R^2 foram calculados para fins comparativos.

4.2.2 Resultados e Discussão

Todos os gráficos presentes nas Figuras 7 a 11, juntos às Tabelas 1 e 2, sintetizam bem o comportamento esperado conforme pesquisas publicadas anteriormente.

Observa-se que, dada a natureza de contagem da variável x_5, o melhor ajuste do modelo de previsão final em todos os casos de proporções simulados ocorreu após a imputação por valores previstos por modelos de Poisson. Apesar desse resultado, imputação por valores gerados de modelos lineares apresentou boa consistência ao comparar com resultados dos modelos de Poisson.

Ao imputar valores pela média e confrontá-los com técnicas de regressão estatística, é possível notar grande diferença no desempenho do modelo final ajustado para y. A Tabela 1, referente aos EQMs de previsão, retrata bem o aumento considerável da variabilidade e das demais estatística relativas à essa técnica, em comparação com as demais. Além disso, ao destacar os valores reais para EQM e R^2 (vide seção 4.1.2), verifica-se que, quanto maior a quantidade de dados faltantes presenta na variável x_5, maior a imprecisão dos valores imputados, independente da técnica implementa.

Ainda assim, a representação das distribuições geradas para os 300 EQMs encontrados para cada proporção (curvas de densidades nas Figuras 7 a 11) retrata boa estabilidade ao imputar valores previstos por modelos de Poisson. Mesmo no caso em que apenas 1% dos valores de x_5 são ausentes (Figura 7), onde é esperada pouca diferença entre técnicas, o destaque da baixa variabilidade dos resultados é nítida ao utilizar regressão de Poisson.

Portanto, de forma expressiva, mesmo em um cenário simples e de pouco impacto, para o conjunto de dados simulados e para todas as proporções de dados faltantes consideradas, os resultados iniciais apontam que imputar valores com uso de técnicas que envolvem modelagem preditiva é mais eficiente, conforme os critérios comparativos escolhidos.

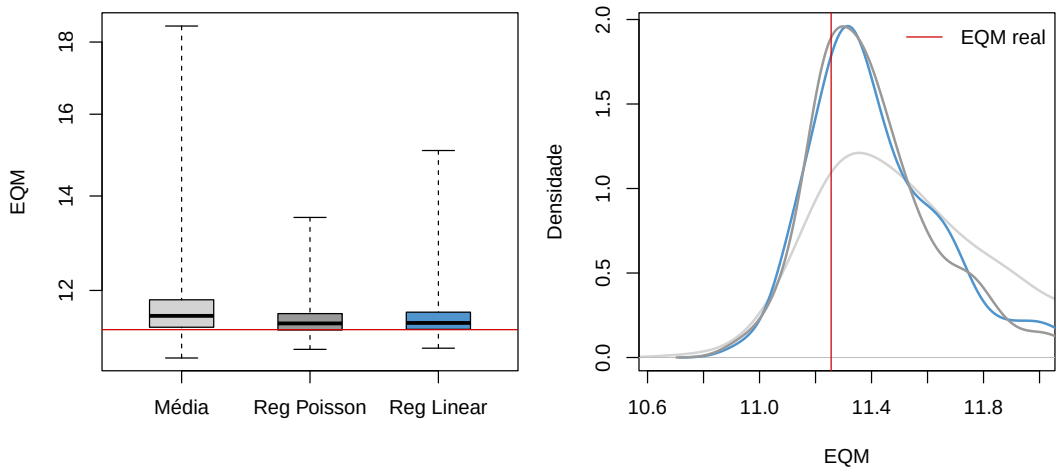

Figura 7: *Boxplots* e curvas de densidade, respectivamente, para EQMs obtidos após implementação de cada técnica de imputação única em cenário simples, com **1%** de dados faltantes.

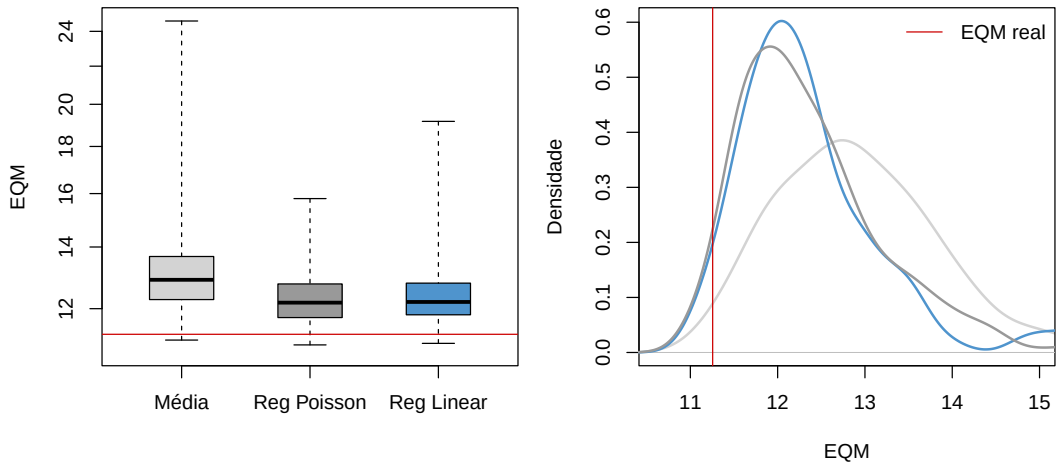

Figura 8: *Boxplots* e curvas de densidade, respectivamente, para EQMs obtidos após implementação de cada técnica de imputação única em cenário simples, com **5%** de dados faltantes.

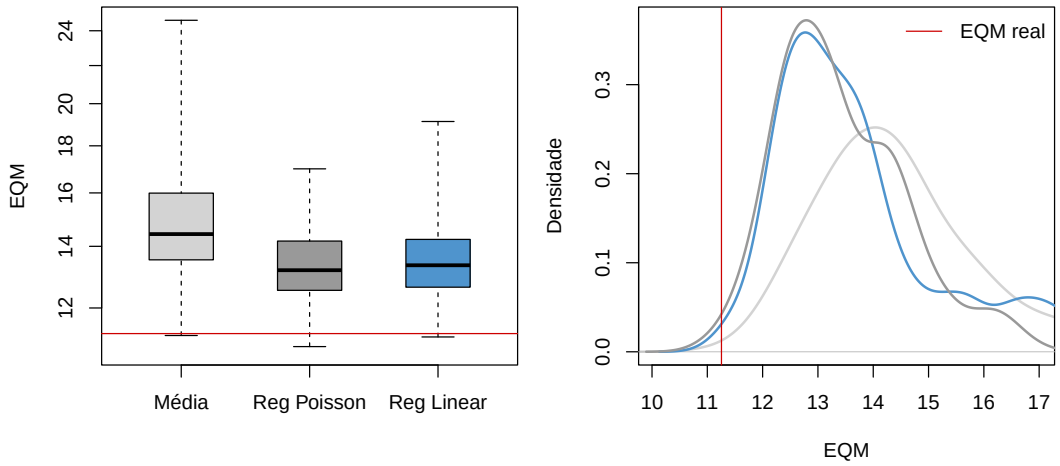

Figura 9: *Boxplots* e curvas de densidade, respectivamente, para EQMs obtidos após implementação de cada técnica de imputação única em cenário simples, com **10%** de dados faltantes.

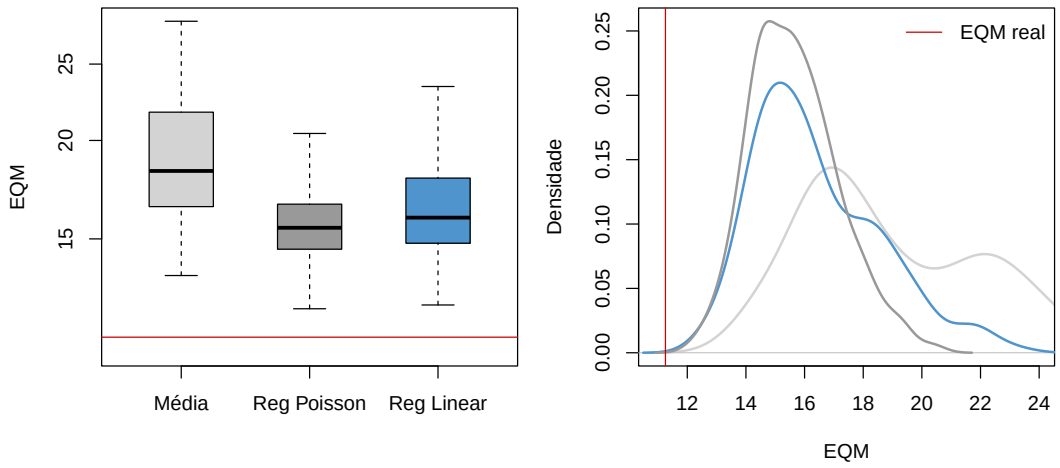

Figura 10: *Boxplots* e curvas de densidade, respectivamente, para EQMs obtidos após implementação de cada técnica de imputação única em cenário simples, com **20%** de dados faltantes.

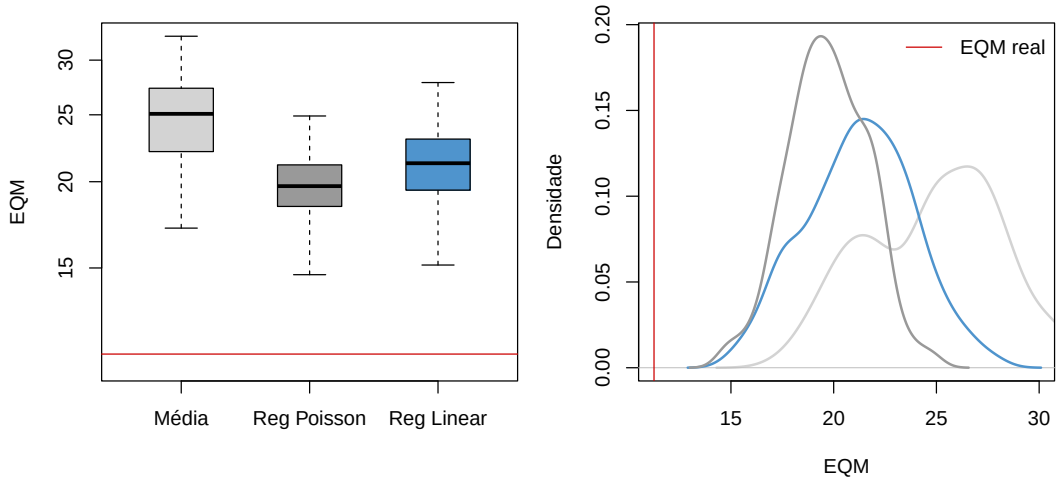

Figura 11: *Boxplots* e curvas de densidade, respectivamente, para EQMs obtidos após implementação de cada técnica de imputação única em cenário simples, com **40%** de dados faltantes.

Tabela 1: Estatísticas descritivas referentes aos **EQMs** obtidos após implementação de técnicas de imputação única em cenário simples, com apenas uma variável com dados faltantes.

Proporção	Técnica	Média	Variância	$Q_{2.5\%}$	Mediana	$Q_{97.5\%}$
1%	Reg Poisson	11.4584	0.1144	11.0502	11.3709	12.4592
	Reg Linear	11.4764	0.2000	11.0776	11.3802	12.3117
	Média	11.6693	0.6208	11.0694	11.5121	12.9461
5%	Reg Poisson	12.3699	0.7133	11.1880	12.1826	14.3774
	Reg Linear	12.5222	1.4882	11.2024	12.2036	15.8249
	Média	13.4183	4.2505	11.3795	12.8979	19.5523
10%	Reg Poisson	13.4164	1.3441	11.6133	13.1882	16.1978
	Reg Linear	13.7589	2.5680	11.6822	13.3525	17.7988
	Média	15.3194	7.1599	12.2178	14.4349	21.6600
20%	Reg Poisson	15.6469	2.2484	13.1358	15.4949	19.1361
	Reg Linear	16.4175	4.6984	13.3116	15.9616	21.7518
	Média	19.0949	10.7023	14.2435	18.2942	26.5414
40%	Reg Poisson	19.7173	3.7306	15.8865	19.7068	23.3552
	Reg Linear	21.2466	6.8146	16.3292	21.2708	26.4561
	Média	24.8313	10.7839	18.7442	25.0792	30.8976

Tabela 2: Estatísticas descritivas referentes aos $\mathbf{R^2}$ obtidos após implementação de técnicas de imputação única em cenário simples, com apenas uma variável com dados faltantes.

Proporção	Técnica	Média	Variância	$Q_{2.5\%}$	Mediana	$Q_{97.5\%}$
1%	Reg Poisson	0.9354	0.0000	0.9297	0.9359	0.9377
	Reg Linear	0.9353	0.0000	0.9306	0.9358	0.9375
	Média	0.9342	0.0000	0.9270	0.9351	0.9376
5%	Reg Poisson	0.9302	0.0000	0.9189	0.9313	0.9369
	Reg Linear	0.9294	0.0000	0.9107	0.9312	0.9368
	Média	0.9243	0.0001	0.8897	0.9272	0.9358
10%	Reg Poisson	0.9243	0.0000	0.9086	0.9256	0.9345
	Reg Linear	0.9224	0.0001	0.8996	0.9247	0.9341
	Média	0.9136	0.0002	0.8778	0.9186	0.9311
20%	Reg Poisson	0.9117	0.0001	0.8921	0.9126	0.9259
	Reg Linear	0.9074	0.0001	0.8773	0.9100	0.9249
	Média	0.8923	0.0003	0.8503	0.8968	0.9197
40%	Reg Poisson	0.8888	0.0001	0.8683	0.8888	0.9104
	Reg Linear	0.8802	0.0002	0.8508	0.8800	0.9079
	Média	0.8599	0.0003	0.8257	0.8585	0.8943

4.3 Parte 2: Imputação Única em Cenários Complexos

De acordo com as discussões propostas na primeira parte, esta segunda parte é inteiramente composta por técnicas de imputação única com uso de modelagem preditiva. Agora, liderada por um cenário mais complexo, duas variáveis apresentam dados faltantes (x_1 e x_5). Conforme análise prévia realizada na parte um, os melhores modelos para obtenção de x_5 consideram x_1 como variável explicativa significativa, e portanto um tratamento mais robusto deve ser considerado para solucionar essa situação. Além disso, a imputação de x_1 também foi direcionado por modelos para variáveis dicotômicas. Assim, três técnicas principais foram utilizadas, sendo elas imputação por valores previstos por: regressão logística dicotômica para x_1; regressão linear para x_5; e de Poisson para x_5. Por fim, alguns resultados sobre geração de modelos desconsiderando uma variável com dados faltantes também foram discutidos nesta etapa.

4.3.1 Método

Neste contexto, a imputação de valores de x_5 depende de outra variável que também apresenta dados faltantes. Para fins ilustrativos, nas situações em que a variável x_1 é covariável nos modelos de imputação de x_5, foi fixada e considerada apenas a proporção de 20% de dados faltantes induzidos em x_1. Dessa forma, dois cenários foram discutidos para exemplificar diferentes contextos de imputação e técnicas utilizadas, que foram comparados posteriormente.

O cenário 1 consiste em desconsiderar a variável x_1, que apresenta dados faltantes, tanto para imputar valores de x_5, quanto para os ajustes dos modelos de previsão finais. Assim, este processo de imputação se baseia apenas em imputar dados de x_5 com base na construção de modelos lineares simples e modelos de Poisson, de modo que a única variável explicativa seja x_2. Por conseguinte, na técnica direcionada por regressão linear, x_5 foi obtido conforme o modelo $X_{5,ij} = \beta_{0j} + \beta_{2j}X_{2i}$. Já para a técnica via regressão de Poisson, o modelo ajustado pôde ser expresso por $X_{5,ij} = \exp(\alpha_j + \beta_{2j}X_{2i})$. Em todas as possíveis equações apresentadas nessa segunda parte, o índice i = 1,2,...,200 se refere à i-ésima observação do conjunto de dados, e j = 1,2,...,300 ao j-ésimo banco de dados simulado.

Por outro lado, no cenário 2 a variável x_1 é indispensável, é utilizada usada para imputar valores de x_5 e para os ajustes dos modelos de previsão finais. Neste caso, primeiro foram imputados valores de x_1 conforme ajustes de modelos de regressão logística dicotômica, e em seguida imputados valores de x_5 com base em modelos lineares e de Poisson.

A escolha das variáveis para os modelos logísticos ocorreu conforme análise de AICs obtidos, sendo aceito que a técnica de imputação para obter x_1 seria modelada pelas covariáveis x_2 e x_4. Apesar da análise da seleção de variáveis indicar x_5 como significativa para obtenção de x_1, neste primeiro momento, como o interesse é trabalhar com apenas técnicas de imputação única, x_5 não foi utilizada. A Equação (4.3.1) exemplifica bem o modelo para probabilidades de 'sucesso' (valor 1) conforme o modelo de regressão logística dicotômica utilizado, que foi o pilar para imputar os valores de x_1:

$$\pi(X_{1,ij}) = \frac{\exp(\beta_{0j} + \beta_{2j}X_{2i} + \beta_{4j}X_{4i})}{1 + \exp(\beta_{0j} + \beta_{2j}X_{2i} + \beta_{4j}X_{4i})}. \tag{4.3.1}$$

Assim, na base de dados j, o valor de x_1 foi imputado como 'sucesso' (valor 1) se $\pi(X_{1,ij}) > 0.5$, e 'fracasso' (valor 0) caso contrário. Em seguida, após os resultados obtidos pela recuperação de todos os 20% de dados faltantes induzidos em x_1, a variável x_5 se torna foco do estudo. Conforme resultados apresentados no primeiro cenário e na seção 4.2, os valores de x_5 foram assim imputados com base na construção de modelos lineares ($X_{5,ij} = \beta_{0j} + \beta_{1j}X_{1,ij} + \beta_{2j}X_{2i}$) e de Poisson ($X_{5,ij} = \exp(\alpha_j + \beta_{1j}X_{1,ij} + \beta_{2j}X_{2i})$), onde as covariáveis explicativas consideradas são x_1 e x_2.

Observa-se que, diferente das Equações (4.2.1) e (4.2.2), a variável $X_{1,ij}$ nestes próximos modelos se refere à nova variável x_1 após o processo de imputação inicial para cada j-ésimo conjunto de dados simulado, em que 20% dos dados foram imputados, e 80% correspondem aos valores originais.

Por fim, os modelos de previsões finais foram ajustados aos novos bancos de dados gerados pelas técnicas de imputação usadas nos cenários 1 e 2. Apenas o EQM de previsão foi usado como métrica para comparar as técnicas de imputação apresentadas em cada cenário, fornecendo informações suficientes para destacar as principais diferenças entre os métodos apresentados.

4.3.2 Resultados e Discussão

Assim como na primeira parte, com base nos EQMs de previsão (Figura 12 e Tabela 3), foi verificado neste segundo momento que, dada a natureza de contagem da variável x_5, o melhor ajuste dos modelos de previsão finais ocorreu após a imputação por valores previstos por modelos de Poisson, em todos os casos de proporções e cenários considerados.

Com respeito ao cenário 1 (variável x_1 é desconsiderada, mesmo com dados faltantes) e à imputação da valores de x_5, é possível notar na Figura 12 que independente das proporções de dados faltantes, os resultados para ambas as técnicas se mostrou muito

próximo, sem destaque significativo de diferença entre as técnicas via modelagem. No entanto, tal fato não se repete no cenário 2, em que a variabilidade dos EQMs obtidos é superior em todas as proporções ao utilizar modelos lineares. Além disso, a Tabela 3 indica que, para a base de dados avaliada, em uma situação de imputação única quanto menor a proporção de dados faltantes, menor a diferença na eficiência das técnicas de imputação com uso de modelagem preditiva, mostrando-se viável nestes casos utilizar técnicas de modelagem mais simples.

A Figura 12 retrata de forma clara a diferença entre os cenários 1 e 2. Ao desconsiderar a variável x_1, que apresenta dados faltantes e é significativa para obter x_5, a perda de desempenho dos modelos preditivos finais após imputação apenas de x_5 é quase certa, mesmo que a variabilidade dos EQMs simulados seja menor para o cenário 1.

No cenário 2 é válido também lembrar que além de imputar valores de x_5 com base em modelos preditivos, dados de x_1 também foram imputados conforme ajustes de regressões logísticas dicotômicas. E, de maneira eficiente, gerou boas aproximações de x_1 para por fim ajustar modelos de imputação de x_5.

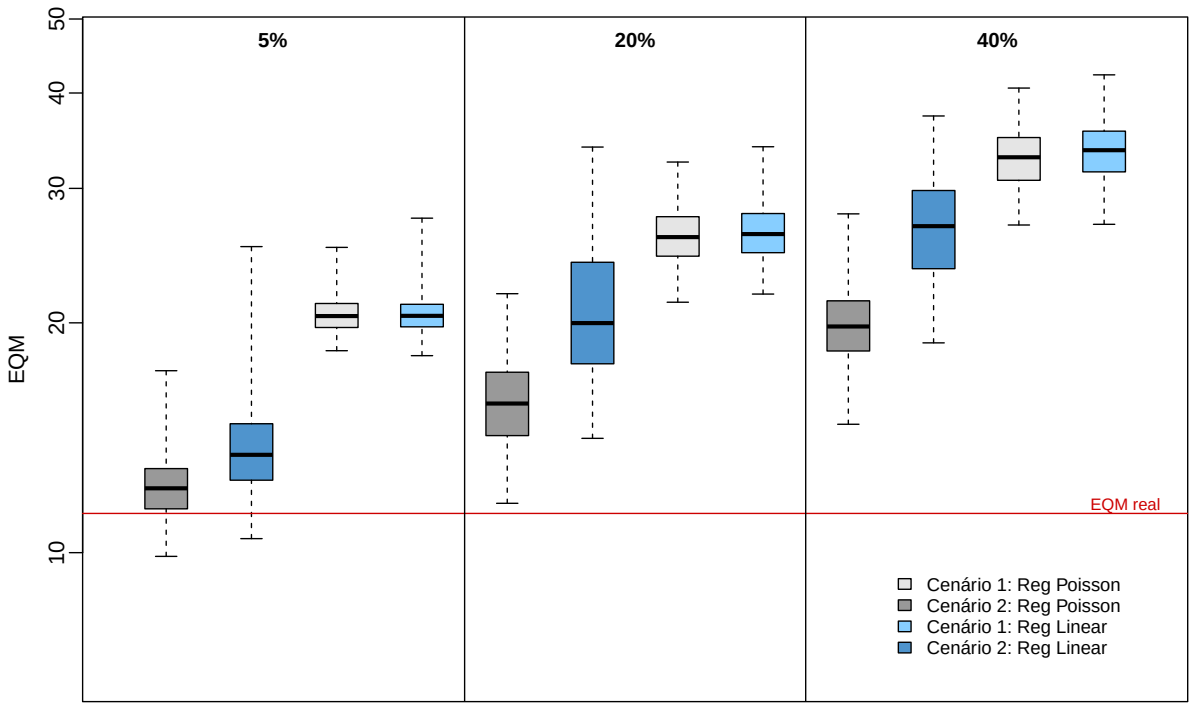

Figura 12: *Boxplots* referentes aos EQMs obtidos nos cenários 1 e 2, conforme cada técnica utilizada para imputar valores de x_5. De acordo também com três proporções de dados faltantes simulados para a variável x_5.

Tabela 3: Estatísticas descritivas referentes aos **EQMs** obtidos após implementação de técnicas de imputação única em cenários complexos (cenários 1 e 2), com mais de uma variável com dados faltantes. Com 20% de dados faltantes fixados em x_1, e representação das proporções de dados faltantes simulados para a variável x_5, e especificação das respectivas técnicas de imputação aplicadas em x_5.

Proporção	Cenário	Técnica	Média	Variância	$Q_{2.5\%}$	Mediana	$Q_{97.5\%}$
1%	2	Reg Poisson	11.4796	0.5097	10.0499	11.4714	12.9261
		Reg Linear	11.7078	1.1444	10.2458	11.5829	13.7977
	1	Reg Poisson	19.0983	0.2289	18.4109	19.0151	20.2073
		Reg Linear	19.1114	0.2864	18.3626	18.9942	20.3558
5%	2	Reg Poisson	12.2827	1.4895	10.4495	12.1422	15.1522
		Reg Linear	14.1028	7.1748	11.0849	13.4324	22.6006
	1	Reg Poisson	20.5981	1.4906	18.8478	20.4128	23.5987
		Reg Linear	20.6466	1.8984	18.8385	20.4257	24.2628
10%	2	Reg Poisson	13.5997	2.4910	11.1866	13.3863	17.1258
		Reg Linear	16.8369	12.6543	12.3722	15.9554	24.7862
	1	Reg Poisson	22.3529	2.7345	19.7372	22.1695	26.3050
		Reg Linear	22.5379	3.2940	19.8984	22.3082	26.9967
20%	2	Reg Poisson	15.8057	4.0059	12.3987	15.6793	19.6121
		Reg Linear	20.8979	16.9086	15.1054	19.9846	30.0155
	1	Reg Poisson	26.0748	4.5929	22.3969	25.8997	30.4276
		Reg Linear	26.4691	5.7600	22.4791	26.1384	32.1497
40%	2	Reg Poisson	20.0127	6.0289	15.5661	19.7810	25.5897
		Reg Linear	26.7316	15.5998	19.5812	26.7709	34.3494
	1	Reg Poisson	32.9800	7.9676	27.9506	32.9509	38.8114
		Reg Linear	33.6992	9.1903	28.2928	33.6617	39.7215

4.4 Parte 3: Imputação Múltipla

Nesta terceira parte foram explorados aspectos de processos de imputação múltipla com uso apenas de modelagem preditiva. Apesar da infinidade de possíveis situações, foram assim exemplificados dois cenários comuns do meio científico e de fácil acesso, voltados à implementação de técnicas de imputação por valores previstos por modelos logísticos dicotômicos para a variável x_1 e lineares para x_5, em um processo iterativo de imputação múltipla, caracterizados por cenários 3 e 4. Por ser tratar de técnicas de imputação múltipla, a construção de bons métodos iterativos foi peça fundamental para gerar bons resultados.

4.4.1 Método

Neste caso, a imputação de valores de x_5 depende de valores imputados de x_1, e o contrário também ocorre, tornando o desenvolvimento iterativo o diferencial. Apesar de inicialmente a imputação ocorrer com base em uma iteração 'individual', o objetivo final está em retratar a eficiência de um método iterativo simultâneo entre as variáveis que apresentam dados faltantes. Além disso, foi fixada a simulação de 20% de dados faltantes em x_1, enquanto que em x_5 foram consideradas todas as demais já trabalhadas anteriormente (1%, 5%, 10%, 20% e 40%).

No cenário 3 a imputação múltipla ocorre de forma individual. Apesar de vários valores para a imputação de uma única observação serem simulados, primeiro valores de x_1 foram imputados com base em regressões logísticas dicotômicas, e em seguida x_5 foi imputado com base em regressões lineares, sendo os valores imputados de x_1 parte de uma das variáveis explicativas para obter x_5. Dessa forma, os modelos de probabilidade das regressões logísticas para imputar x_1 foram gerados como na Equação (4.3.1) e, com respeito às técnicas de imputação para x_5, no cenário 3 os modelos lineares foram simplificados por $X_{5,ij} = \beta_{0j} + \beta_{1j}X_{1,ij} + \beta_{2j}X_{2i}$, sendo as mesmas expressões numéricas utilizadas no cenário 2. Válido também citar que em todas as possíveis equações apresentadas nessa terceira parte o índice i = 1,2,...,200 se refere à i-ésima observação do conjunto de dados, e j = 1,2,...,300 ao j-ésimo banco de dados simulado.

Já no cenário 4 a imputação múltipla ocorre de forma simultânea e sequencial entre as variáveis, tornando o sistema ainda mais complexo. A várias imputações são realizadas até que a convergência seja alcançada e todos os valores ausentes sejam imputados com base em critérios estatísticos. Este processo também é conhecido por imputação multivariada simultânea. Com base em técnicas de modelagem preditiva, a imputação de x_1 se deu por modelos de regressão logística dicotômica, com x_2, x_4 e x_5 como covariáveis explicativas. Já a imputação de x_5 se deu por modelos lineares, sendo x_1 e x_2 as covariáveis explicativas, com as previsões de ambos os modelos correlacionadas. O Exemplo 1, representado pela Tabela 4, retrata um passo a passo simples para o mesmo método de imputação realizado no cenário 4, porém em um conjunto de dados fictícios.

Exemplo 1. Suponha uma base de dados com variáveis com as mesmas características das utilizadas no trabalho, porém com apenas 10 observações no total. Considere também os dados faltantes da i-ésima observação de x_1 representados por $\mathrm{NA}_{1,i}$, e de x_5 por $\mathrm{NA}_{5,i}$, $i = 1, 2..., 10$. Neste exemplo e no cenário 4 a ordem de prioridade de imputação das variáveis ocorre conforme: (1) ordem crescente (da primeira $i = 1$ à última observação $i = 10$); (2) caso apenas uma das variáveis apresente valor ausente na observação i, o valor é obrigatoriamente imputado para esta variável; (3) e caso x_1 e x_5 apresentem

dados faltantes na mesma observação, imputa-se primeiro x_1, em seguida x_5.

Tabela 4: Base de dados referente ao Exemplo 1.

i	$\mathbf{x_1}$	$\mathbf{x_5}$	$\mathbf{x_4}$	$\mathbf{x_3}$	$\mathbf{x_2}$
1	$\text{NA}_{1,1}$	$x_{5,1}$	$x_{4,1}$	$x_{3,1}$	$x_{2,1}$
2	$x_{1,2}$	$x_{5,2}$	$x_{4,2}$	$x_{3,2}$	$x_{2,2}$
3	$\text{NA}_{1,3}$	$\text{NA}_{5,3}$	$x_{4,3}$	$x_{3,3}$	$x_{2,3}$
4	$x_{1,4}$	$x_{5,4}$	$x_{4,4}$	$x_{3,4}$	$x_{2,4}$
5	$x_{1,5}$	$x_{5,5}$	$x_{4,5}$	$x_{3,5}$	$x_{2,5}$
6	$x_{1,6}$	$\text{NA}_{5,6}$	$x_{4,6}$	$x_{3,6}$	$x_{2,6}$
7	$\text{NA}_{1,7}$	$x_{5,7}$	$x_{4,7}$	$x_{3,7}$	$x_{2,7}$
8	$x_{1,8}$	$x_{5,8}$	$x_{4,8}$	$x_{3,8}$	$x_{2,8}$
9	$x_{1,9}$	$x_{5,9}$	$x_{4,9}$	$x_{3,9}$	$x_{2,9}$
10	$x_{1,10}$	$\text{NA}_{5,10}$	$x_{4,10}$	$x_{3,10}$	$x_{2,10}$

Tendo em vista a Tabela 4 como exemplo para representação e posição dos dados faltantes, o passo a passo para imputação múltipla simultânea neste cenário pode ser simplificado por:

1. Imputa-se $\mathbf{NA_{1,1}}$ de acordo com a previsão do modelo logístico adequado, com auxílio das covariáveis significativas restantes cujas observações são completas. Neste primeiro momento as demais observações ainda não são consideradas por causa dos dados faltantes. Qualifica-se então i = 2,4,5,8,9 para ajuste do modelo.

2. Imputa-se $\mathbf{NA_{1,3}}$ de acordo com a previsão do modelo logístico adequado, com uso das covariáveis significativas restantes cujas observações são completas e considerando acréscimo de valores imputados em $i = 1$ para x_1. Qualifica-se então i = 1,2,4,5,8,9 para ajuste do modelo. As demais observações ainda não são consideradas por causa dos dados faltantes.

3. Imputa-se $\mathbf{NA_{5,3}}$ de acordo com a previsão do modelo linear adequado, com uso das covariáveis significativas restantes cujas observações são completas, considerando acréscimo de valores imputados em $i = 1$ para x_1. Qualifica-se então i = 1,2,4,5,8,9 para ajuste do modelo. As demais observações ainda não são consideradas por causa dos dados faltantes.

4. Imputa-se $\mathbf{NA_{5,6}}$ de acordo com a previsão do modelo linear adequado, com uso das covariáveis significativas restantes cujas observações são completas, considerando acréscimo de valores imputados em $i = 1$ e 3, para x_1 e x_5. Qualifica-se então i = 1,2,3,4,5,8,9 para ajuste do modelo. As demais observações ainda não são consideradas por causa dos dados faltantes.

5. Imputa-se $\mathbf{NA_{1,7}}$ de acordo com a previsão do modelo logístico adequado, com uso das covariáveis significativas restantes cujas observações são completas, considerando acréscimo de valores imputados em i = 1,3 e 6, para x_1 e x_5. Qualifica-se então i = 1,2,3,4,5,6,8,9 para ajuste do modelo. As demais observações ainda não são consideradas por causa dos dados faltantes.

6. Imputa-se $\mathbf{NA_{5,10}}$ de acordo com a previsão do modelo linear adequado, com uso das covariáveis significativas restantes cujas observações são completas, considerando acréscimo de valores imputados em i = 1,3 e 6, para x_1 e x_5. Logo, com exceção de i = 10, todas as demais observações são qualificadas para ajustar o modelo de imputação.

7. Após finalizar o processo da primeira imputação completa da base de dados (imputação 1), retorna-se ao primeiro dado faltante $\mathbf{NA_{1,1}}$ para iniciar um novo processo (imputação 2). Assim, é necessário reiniciar com imputação de todos os $\mathbf{NA_{1,i}}$ e $\mathbf{NA_{5,i}}$ seguindo a mesma ordem de imputação guiada nos passos 1 a 6 referentes à imputação 1, e seus respectivos modelos. Agora, a imputação 2 ocorre com auxílio de todas as covariáveis significativas do modelo considerando os dados imputados anteriormente na imputação 1. Deste momento em diante, com exceção da observação que será imputada, todas as demais são qualificadas para ajustar o modelo de imputação.

8. O processo realizado em todas as próximas imputações deve ser semelhante à imputação 2. Por consequência, a imputação m é feita com base na imputação $m-1$, em que m = 1,2,...,M representa o momento de cada imputação. Após o total de M imputações, seguindo a convergência, a base final com os dados imputados será definida e pronta para análise.

Com os modelos das técnicas de imputação já definidos, os resultados dos métodos iterativos dos cenários 3 e 4 ocorreram com auxílio da função **mice()** do pacote MICE no R. Assim, em ambos os cenários, conforme sugestões de Bodner (2008) e Burren (2018), a decisão da quantidade máxima M de imputações múltiplas para cada variável ocorreu com base na proporção de dados faltantes. Para x_1, que apresentou sempre 20% de dados ausentes, foi fixado M = 12 imputações. Já para x_5, devido às cinco proporções consideradas na simulação, foi estabelecida a quantidade de M = 24 imputações para todos os casos, devido a questões relacionadas ao elevado tempo de execução da simulação e baixa complexidade dos dados (distribuição uniforme dos dados faltantes).

Por último, os modelos de previsões finais foram ajustados aos novos bancos de dados gerados pelas técnicas de imputação usadas nos cenários 3 e 4. Assim como na segunda parte, apenas o EQM de previsão foi usado como métrica para comparar

as técnicas de imputação apresentadas em cada cenário da terceira parte, fornecendo informações que destacam as principais diferenças entre os métodos apresentados. As comparações finais realizadas ocorreram com base nos critérios estabelecidos na seção 4.1.4, referentes principalmente às técnicas de ajuste de modelos de regressão linear para imputar a valores de x_5, comparando os cenários 1, 2, 3 e 4, discutidos nas partes 2 e 3 deste trabalho.

4.4.2 Resultados e Discussão

Frente aos EQMs de previsão obtidos, a Figura 13 deixa clara a diferença do uso de imputação em x_5 por valores previstos por regressões lineares entre os quatro cenários apresentados nas partes 2 e 3. Neste contexto com apenas uso de modelos lineares, o destaque maior de desempenho foi nas técnicas implementadas no cenário 4, utilizando a imputação múltipla de forma simultânea e sequencial entre as variáveis que apresentaram dados faltantes. Também foram observados indícios de que quanto maior a proporção de dados faltantes, melhor o desempenho de técnicas de imputação múltipla quando comparadas com as demais (Tabela 5).

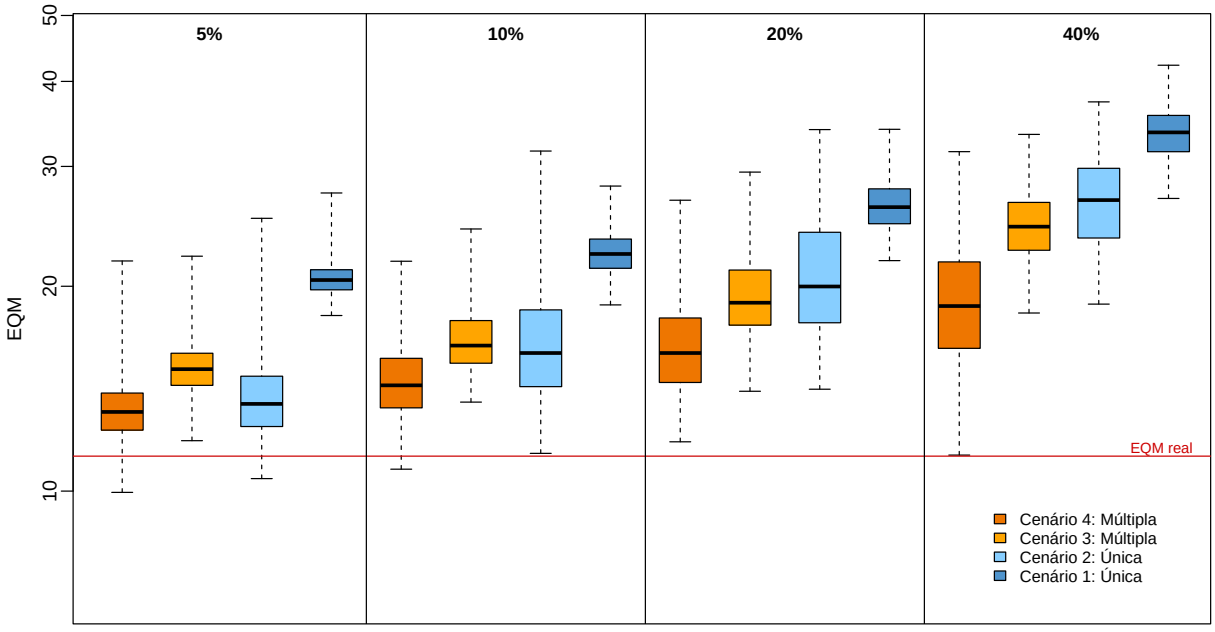

Figura 13: *Boxplots* referentes aos EQMs obtidos nos cenários 1 a 4, para representação de resultados de técnicas com uso, apenas, de modelos de **regressão linear** para imputar valores de x_5. Com representação de quatro proporções de dados faltantes simulados para a variável x_5.

Observa-se na Figura 13 que, ao comparar os cenários 2 e 4 em situações nas quais x_5 apresenta baixas proporções de dados faltantes (desconsiderando a variabilidade dos resultados), parece haver pouca diferença significativa entre as técnicas implementadas. Válido lembrar que os cenários 2 e 4 representados na Figura 13 correspondem aos EQMs

obtidos após 20% de dados faltantes imputados em x_1 com uso de técnicas de imputação via modelos logísticos, com dados faltantes de x_5 imputados por modelos lineares.

Aparentemente, para proporções menores de dados faltantes, ao utilizar apenas regressões lineares, o cenário 2 de imputação única propôs melhores resultados que o cenário 3 de imputação múltipla. No entanto, neste contexto a variabilidade dos EQMs obtidos no cenário 3 se mostrou muito inferior para todas as proporções de dados faltantes, e o desempenho do cenário 3 para 20% e 40% de dados faltantes em x_5 foi superior quando comparado com o cenário 2. Além do mais, novamente o pior desempenho apresentado ocorreu no cenário 1, em que a variável x_1 (com dados faltantes) é desconsiderada e não tratada.

Ao comparar resultados após uso de regressões de Poisson nos cenários 1 e 2 (imputação única), com os cenários 3 e 4 (imputação múltipla), algumas diferenças contraintuitivas foram verificadas.

Em contextos em que a proporção de dados faltantes é muito baixa para a variável x_5 e alta para a variável x_1 (20% de dados faltantes) (Figura 14), a imputação única com uso de regressão de Poisson para valores de x_5 e imputação única com uso de regressão logística em x_1 (cenário 2) se mostrou a mais eficiente. Todavia, ainda referente ao resultados de baixas proporções, as curvas de densidade dos EQMs encontrados apontam fatores a favor do cenário 4, ao indicar boa simetria e baixa dispersão dos dados, sendo muito semelhante à distribuição dos EQMs encontrados no cenário 2.

Além do mais, à medida que a proporção de dados faltantes para a variável x_5 aumenta e continua alta para a variável x_1, com 20% de dados faltantes (Figura 15 e Tabela 5), o processo realizado no cenário 2 de imputação única por modelos de Poisson acaba perdendo desempenho quando comparado ao cenário 4. Contudo, neste contexto de aumento das proporções, as distribuições presentes na Figura 15 indicam melhor estabilidade e simetria para os cenários 1,2 e 3, e a presença de alguns valores de EQMs atípicos que influenciaram os resultados do cenário 4.

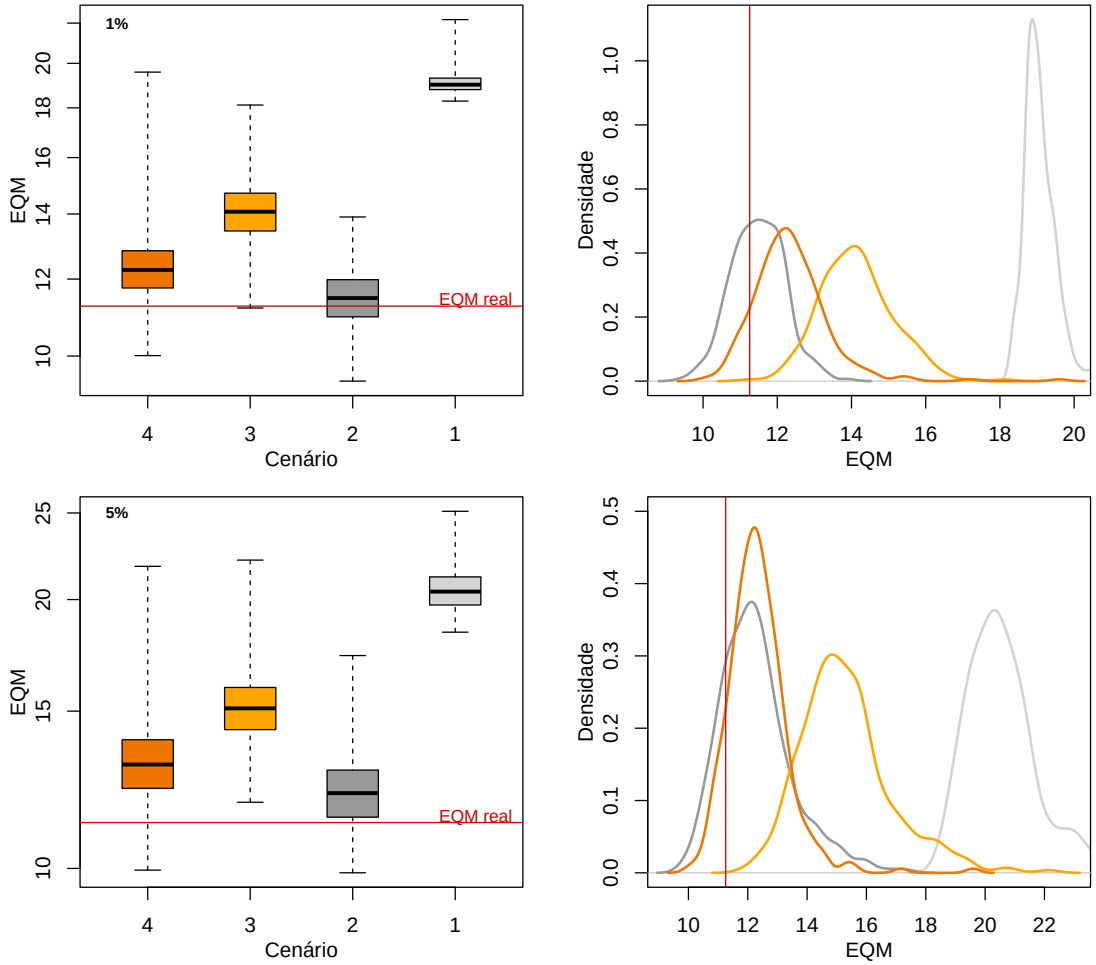

Figura 14: *Boxplots* e curvas de densidade, respectivamente, para EQMs obtidos após implementação de cada técnica: imputação única com modelos de Poisson (cenários 1 e 2); e imputação múltipla com modelos lineares (cenários 3 e 4). De acordo com **1%** e **5%** de dados faltantes induzidos em x_5, respectivamente.

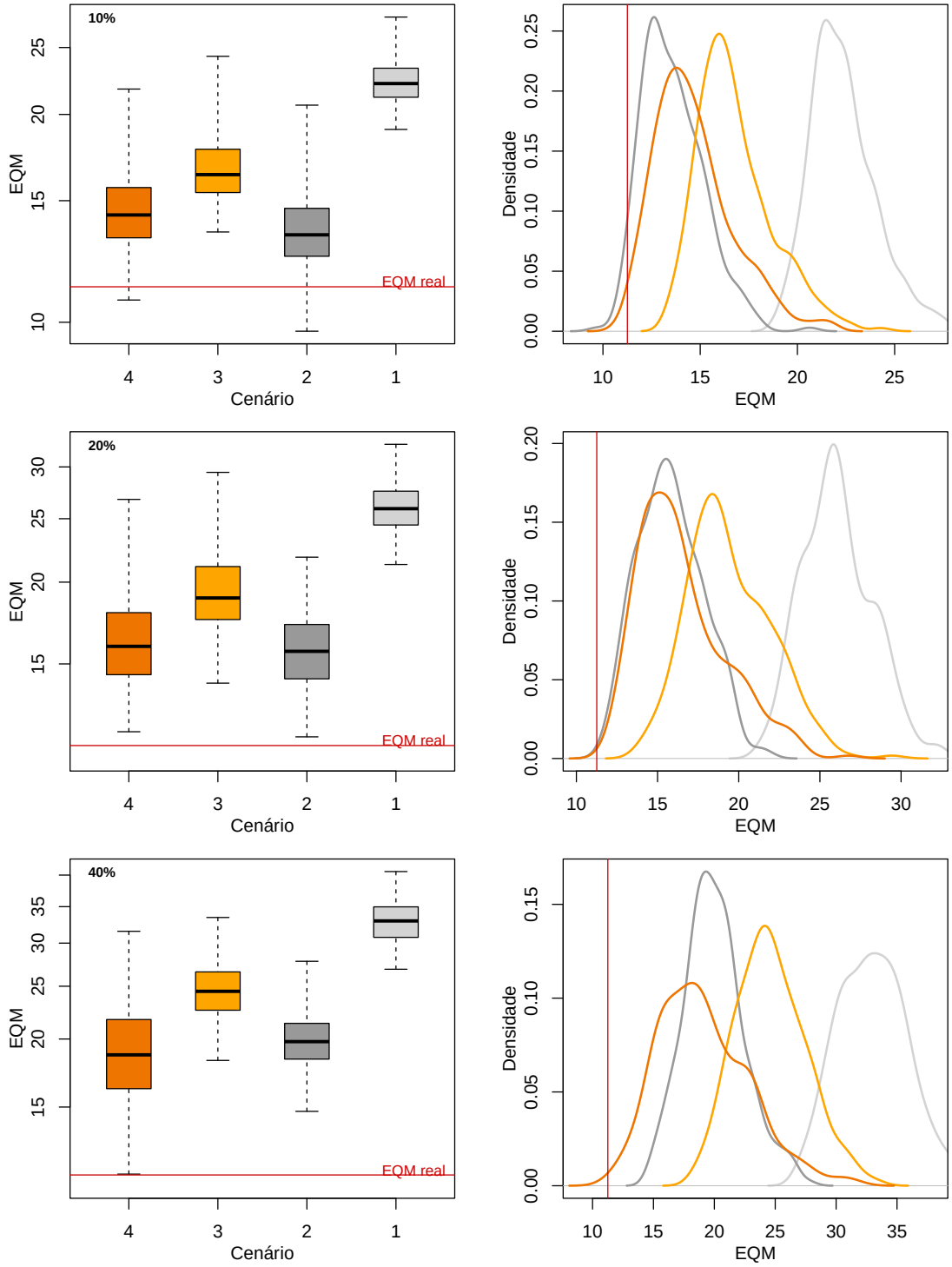

Figura 15: *Boxplots* e curvas de densidade, respectivamente, para EQMs obtidos após implementação de cada técnica: imputação única com modelos de Poisson (cenários 1 e 2); e imputação múltipla com modelos lineares (cenários 3 e 4). De acordo com **10%**, **20%** e **40%** de dados faltantes induzidos em x_5, respectivamente.

Tabela 5: Estatísticas descritivas referentes aos **EQMs** obtidos após implementação de técnicas de imputação única (cenários 1 e 2) e múltipla (cenários 3 e 4), com mais de uma variável com dados faltantes. Com 20% de dados faltantes fixados em x_1, representação das proporções de dados faltantes simulados para a variável x_5, e especificação da técnica de imputação aplicada em x_5.

Proporção	Cenário	Imputação	Técnica	Média	Variância	$Q_{2.5\%}$	Mediana	$Q_{97.5\%}$
1%	4	Múltipla	Reg Linear	12.345	1.037	10.776	12.265	14.484
	3	Múltipla	Reg Linear	14.129	0.972	12.403	14.074	16.140
	2	Única	Reg Linear	11.708	1.144	10.246	11.583	13.798
	1	Única	Reg Linear	19.111	0.286	18.363	18.994	20.356
	2	Única	Reg Poisson	11.480	0.510	10.050	11.471	12.926
	1	Única	Reg Poisson	19.098	0.229	18.411	19.015	20.207
5%	4	Múltipla	Reg Linear	13.323	2.674	10.954	13.072	17.480
	3	Múltipla	Reg Linear	15.301	2.391	12.899	15.108	19.031
	2	Única	Reg Linear	14.103	7.175	11.085	13.432	22.601
	1	Única	Reg Linear	20.647	1.898	18.839	20.426	24.263
	2	Única	Reg Poisson	12.283	1.489	10.450	12.142	15.152
	1	Única	Reg Poisson	20.598	1.491	18.848	20.413	23.599
10%	4	Múltipla	Reg Linear	14.646	4.128	11.614	14.303	19.181
	3	Múltipla	Reg Linear	16.758	3.617	13.895	16.363	21.109
	2	Única	Reg Linear	16.837	12.654	12.372	15.955	24.786
	1	Única	Reg Linear	22.538	3.294	19.898	22.308	26.997
	2	Única	Reg Poisson	13.600	2.491	11.187	13.386	17.126
	1	Única	Reg Poisson	22.353	2.735	19.737	22.169	26.305
20%	4	Múltipla	Reg Linear	16.468	7.155	12.775	15.956	22.911
	3	Múltipla	Reg Linear	19.421	6.505	14.985	18.917	24.780
	2	Única	Reg Linear	20.898	16.909	15.105	19.985	30.016
	1	Única	Reg Linear	26.469	5.760	22.479	26.138	32.150
	2	Única	Reg Poisson	15.806	4.006	12.399	15.679	19.612
	1	Única	Reg Poisson	26.075	4.593	22.397	25.900	30.428
40%	4	Múltipla	Reg Linear	19.112	13.803	13.151	18.705	27.452
	3	Múltipla	Reg Linear	24.645	8.221	19.738	24.469	30.775
	2	Única	Reg Linear	26.732	15.600	19.581	26.771	34.349
	1	Única	Reg Linear	33.699	9.190	28.293	33.662	39.721
	2	Única	Reg Poisson	20.013	6.029	15.566	19.781	25.590
	1	Única	Reg Poisson	32.980	7.968	27.951	32.951	38.811

5 Guia Prático para Imputação

5.1 Tipo de Imputação

Fundamentado-se pelos resultados obtidos no Capítulo 4 e na literatura, é possível traçar um guia prático e simples para decisão de cada técnica de imputação para ajustes de modelos de regressão. Harrell *et al.* (2015) menciona, por linhas gerais, ser possível decidir cada método de imputação com base na proporção de dados faltantes p de qualquer uma das variáveis incompletas, e na característica da variável. Assim, no cenário de imputação para gerar modelos preditivos, é válido considerar os seguintes pontos:

1. $p < 3\%$: não importa muito a forma como se imputam os valores faltantes ou se ajusta a variância das estimativas do coeficiente de regressão para ter os dados imputados. Neste caso, é provável ser suficiente usar técnicas de imputação única para obter bons resultados. Para variáveis quantitativas, a imputação de dados faltantes com o valor mediano dos dados restantes pode funcionar bem; para variáveis qualitativas, pode ser utilizada a categoria mais frequente. Apesar disso, os resultados se tornam ainda mais robustos ao considerar técnicas que envolvem modelagem preditiva. A análise completa dos casos é também uma opção nesta sitaução, assim a imputação múltipla pode ser necessária para verificar se a abordagem simples de imputação única 'funcionou'.

2. $p \geq 3\%$: recomenda-se utilizar técnicas de imputação múltipla com um número de imputações M igual à $\max\{5\ ,\ 100p\}$, embora amostras muito grandes possam permitir um número menor de imputações. A imputação deve ser levada em conta na estimativa da matriz de covariância para as estimativas dos parâmetros finais. Se possível, é sugerido utilizar a distribuição t de Student ao invés da distribuição Gaussiana para os testes e intervalos de confiança.

3. Em situações com muitas variáveis explicativas com dados faltantes e significativas, deve-se considerar as mesmas observações acima, porém a prioridade neste caso precisa ser por uso de técnicas de imputação múltipla, uma vez que mais imputações serão requeridas. Neste caso a ordem em que as variáveis são imputadas acaba sendo importante. Segundo Harrell *et al.* (2015), a decisão sobre a ordem de imputação das variáveis pode ser concluída após simular múltiplas imputações utilizando diferentes ordens de variáveis, sendo possivelmente mais eficiente iniciar a imputação pela variável com maior número de dados faltantes, para que a inicialização de outras variáveis incompletas para medianas tenha menos impacto.

Em contextos mais gerais de imputação, sem necessariamente envolver regressão

estatística, Harrell *et al.* (2001) e Nunes (2007) simplificam a decisão sobre imputação única ou múltipla da seguinte forma, se: $p \leq 0.05$, usar imputação única ou apenas analisar os dados completos; $0.05 < p < 0.15$, imputação única pode ser utilizada, porém é mais indicado uso da múltipla; e $p \geq 0.15$, a imputação múltipla é recomendada na maioria dos casos. Além disso, é certo mencionar que a avaliação do motivo da ausência de informação, na maior parte dos casos, é muito mais importante que a quantidade de dados faltantes. Com isso, as observações apresentadas na seção 3.1 devem ser avaliadas antes da implementação de cada técnica.

5.2 Seleção de Variáveis e Processos

Em técnicas que envolvem uso de modelagem preditiva, a decisão sobre as variáveis que serão utilizadas para imputar dados faltantes de outra variável se torna ponto crucial.

Apesar de estudos relatarem que quanto maior o uso de informações melhor os resultados da imputação, nem sempre a maior quantidade de variáveis explicativas implica em bons valores imputados. Buuren (2018) menciona que, para cenários de imputação, é interessante selecionar subconjunto de covariáveis que não apresentam mais do que 15 a 25 variáveis. Em complemento aos resultados prévios da literatura, e conclusões deste trabalho, é minimamente recomendável considerar os seguintes pontos para seleção das covariáveis usadas em técnicas de imputação por modelagem estatística:

1. Avaliar os fatores que causaram a presença de dados faltantes em cada covariável;
2. Inspecionar se alguma outra variável do banco de dados apresenta certa influência nos dados faltantes analisados;
3. Especificar a natureza da variável incompleta que será imputada, e verificar sua quantidade de dados faltantes;
4. Decidir sobre o tipo de imputação: única ou múltipla;
5. Selecionar o tipo regressão com base na natureza da variável;
6. Inicialmente, incluir todas covariáveis possíveis que aparecem nos modelos ajustados aos dados completos;
7. Fazer análises de diagnóstico e selecionar as variáveis conforme medidas de comparação entre modelos, até obter o melhor ajuste possível para os dados completos.

Quando um modelo é superajustado, ou seja, quando inclui preditores redundantes, é possível que haja uma redução na precisão das estimativas finais, mas isso não deve

causar problemas como viés. Por outro lado, se preditores importantes forem omitidos, pode haver viés. Portanto, é preferível errar para mais, ou seja, superajustar, do que para menos (KENWARD; CARPENTER, 2007).

5.3 Pacotes e Funções no R

Com o objetivo de fornecer orientação, tem-se uma breve introdução e sugestão de alguns pacotes e funções no R (versão 4.3.2) para trabalhar com imputação de dados. Todas as informações a seguir foram baseadas em Buuren (2018) e Harrell *et al.* (2015), junto a documentos oficiais do software R.

A função **mice()** do pacote MICE permite ajustar cenários de imputação múltipla de dados faltantes, ao considerar um conjunto de preditores e retornar uma única imputação para cada entrada em falta na variável incompleta, de tal forma que antes de imputar cada dado faltante todos os outros valores gerados para a mesma variável são avaliados e selecionados pela função de maneira otimizada. Além disso o algoritmo de **mice()** é capaz de preencher dados faltantes em bases que contenham uma mistura de variáveis contínuas, binárias, categóricas não ordenadas e categóricas ordenadas. A função também possibilita lidar com dados contínuos de dois níveis e manter a consistência entre as imputações por meio de imputação passiva. Por fim, consegue gerar vários gráficos de diagnóstico para inspecionar a qualidade das imputações (BUUREN; GROOTHUIS-OUDSHOORN, 2023).

Semelhante ao pacote MICE, o pacote Hmisc apresenta funções interessantes, como **aregImpute()**. Com essa função é possível imputar valores de forma eficiente e paralela, o que pode ser útil para cenários de grandes bancos de dados. Em especial é comumente utilizada para imputação múltipla com uso de regressão aditiva, bootstrapping e correspondência de média preditiva (JR, 2023).

Outra função conhecida, a **impute()** do pacte mlr, realiza a imputação em um conjunto de dados e retorna, juntamente com a nova base de dados imputados, um objeto 'ImputationDesc' que pode conter coeficientes 'aprendidos' e dados importantes. Com esse objeto gerado, é possível realizar um processo de reimputação com um novo conjunto de dados obtidos. As técnicas de imputação podem ser especificadas de acordo com determinadas características ou classes de características desejadas, sendo possível fornecer um objeto arbitrário, usar um método de imputação integrado ou criar um método personalizado (BISCHL *et al.*, 2016).

Adicionalmente, para situações ainda mais especificas de estudo, missForest é um pacote que usa florestas aleatórias para imputação de dados (STEKHOVEN; BÜHLMANN, 2012). Enquanto o Amelia é um pacote que usa modelos de equações estruturais

para imputação de dados (HONAKER; KING; BLACKWELL, 2011).

6 Conclusão e Considerações

O objetivo principal deste trabalho foi retratar possíveis diferenças entre técnicas de imputação em alguns cenários de bancos de dados incompletos, com foco maior em métodos que envolvem modelagem preditiva. Apesar da literatura apresentar uma infinidade de métodos de imputação, o tema no Brasil não é tão explorado na literatura, por consequência, desconsiderado em muitas pesquisas. Além do assunto ser pouco divulgado, mesmo quando conhecido há dificuldade por partes de analistas em implementar técnicas de imputação com uso de regressão estatística, por pouca familiaridade com o tema.

Dessa forma, pontos contra e a favor da aplicação de técnicas de imputação foram avaliados, com retrato do nível de dificuldade de cada método. Alguns resultados obtidos pelo estudo de simulação deste trabalho mostraram grandes vantagens ao se utilizar técnicas de regressão estatística para imputar dados faltantes.

Os melhores resultados foram obtidos nos processos que envolveram imputação múltipla, em especial após valores imputados de forma simultânea e sequencial entre as variáveis incompletas. Ainda assim, em alguns cenários de baixa proporção de dados faltantes, alguns resultados de imputações únicas se mostraram ainda mais eficientes quando comparados com de imputações múltiplas, devido em principal à natureza da variável.

Como esperado, situações simuladas que apresentaram mais de uma variável com dados faltantes, significativas para o ajuste de um modelo final, apresentaram maior complexidade e maior imprecisão para os erros quadráticos médios (EQM) encontrados, quando comparadas com casos de apenas uma variável incompleta. Também foi observado que, em todos os cenários analisados, desconsiderar observações ou até variáveis com dados faltantes gerou os piores resultados para o ajuste do modelo final de comparação.

Portanto, foram notados três principais fatores de influência no desempenho das técnicas: quantidade de dados faltantes, número de variáveis incompletas e natureza de tais variáveis. E assim, com base na literatura e nos resultados deste trabalho, um guia prático foi documentado, a fim de auxiliar pesquisadores quanto ao tratamento de bases incompletas.

Sugere-se para trabalhos futuros relacionados à imputação explorar novos cenários de dados faltantes, e propor algoritmos ou guias de solução cada vez mais especializados, com a finalidade de habituar a comunidade científica quanto ao tema e melhorar a qualidade das informações geradas por análises de conjunto de dados incompletos.

Referências

AGRESTI, A. *An introduction to categorical data analysis*. 2nd. ed. [S.l.]: John Wiley & Sons, 2007. 372 p.

AGRESTI, A. *An introduction to categorical data analysis*. [S.l.]: John Wiley & Sons, 2018.

AMBLER, G.; OMAR, R. Z.; ROYSTON, P. A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. *Statistical methods in medical research*, Sage Publications Sage UK: London, England, v. 16, n. 3, p. 277–298, 2007.

BISCHL, B.; LANG, M.; KOTTHOFF, L.; SCHIFFNER, J.; RICHTER, J.; STUDERUS, E.; CASALICCHIO, G.; JONES, Z. M. mlr: Machine learning in r. *Journal of Machine Learning Research*, v. 17, n. 170, p. 1–5, 2016. Disponível em: ⟨https://jmlr.org/papers/v17/15-066.html⟩.

BODNER, T. E. What improves with increased missing data imputations? *Structural equation modeling: a multidisciplinary journal*, Taylor & Francis, v. 15, n. 4, p. 651–675, 2008.

BUCK, S. F. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 22, n. 2, p. 302–306, 1960.

BUUREN, S. V. *Flexible imputation of missing data*. [S.l.]: CRC press, 2018.

BUUREN, S. van; GROOTHUIS-OUDSHOORN, K. *mice function*. 2023. ⟨https://www.rdocumentation.org/packages/mice/versions/3.16.0/topics/mice⟩. Acessado em 27 de novembro, 2023.

CELLA, L. O. G. Regressão ordinal bayesiana. Dissertação (Mestrado em Estatística), Instituto de Ciências Exatas, Departamento de Estatística, Universidade de Brasília, nov. 2013.

CONSUL, P.; FAMOYE, F. Generalized poisson regression model. *Communications in Statistics-Theory and Methods*, Taylor & Francis, v. 21, n. 1, p. 89–109, 1992.

DRAPER, N. R.; SMITH, H. *Applied regression analysis*. [S.l.]: John Wiley & Sons, 1998. v. 326.

ENDERS, C. K. *Applied missing data analysis*. [S.l.]: Guilford press, 2010.

ENGELS, J. M.; DIEHR, P. Imputation of missing longitudinal data: a comparison of methods. *Journal of clinical epidemiology*, Elsevier, v. 56, n. 10, p. 968–976, 2003.

HARRELL, F. E. *et al. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. [S.l.]: Springer, 2001. v. 608.

HARRELL, F. E. *et al. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. [S.l.]: Springer, 2015. v. 3.

HEIJDEN, G. J. Van der; DONDERS, A. R. T.; STIJNEN, T.; MOONS, K. G. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. *Journal of clinical epidemiology*, Elsevier, v. 59, n. 10, p. 1102–1109, 2006.

HONAKER, J.; KING, G.; BLACKWELL, M. Amelia II: A program for missing data. *Journal of Statistical Software*, v. 45, n. 7, p. 1–47, 2011.

JR, F. E. H. *aregImpute function*. 2023. ⟨https://www.rdocumentation.org/packages/Hmisc/versions/5.1-1/topics/aregImpute⟩. Acessado em 27 de novembro, 2023.

KENWARD, M. G.; CARPENTER, J. Multiple imputation: current perspectives. *Statistical methods in medical research*, Sage Publications Sage UK: London, England, v. 16, n. 3, p. 199–218, 2007.

KUTNER, M. H.; NACHTSHEIM, C. J.; NETER, J.; LI, W. *et al. Applied linear statistical models*. [S.l.]: McGraw-Hill New York, 2005.

MCKNIGHT, P. E.; MCKNIGHT, K. M.; SIDANI, S.; FIGUEREDO, A. J. *Missing data: A gentle introduction*. [S.l.]: Guilford Press, 2007.

NUNES, L. N. Métodos de imputação de dados aplicados na área da saúde. 2007.

NUNES, L. N.; KLÜCK, M. M.; FACHEL, J. M. G. Comparação de métodos de imputação única e múltipla usando como exemplo um modelo de risco para mortalidade cirúrgica. *Revista Brasileira de Epidemiologia*, SciELO Public Health, v. 13, p. 596–606, 2010.

PEUGH, J. L.; ENDERS, C. K. Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of educational research*, Sage Publications Sage CA: Thousand Oaks, CA, v. 74, n. 4, p. 525–556, 2004.

PROVOST, F.; FAWCETT, T. Data science and its relationship to big data and data-driven decision making. *Big data*, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 1, n. 1, p. 51–59, 2013.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2023. Disponível em: ⟨https://www.R-project.org/⟩.

RODRIGUES, S. C. A. *Modelo de regressão linear e suas aplicações*. Tese (Doutorado) — Universidade da Beira Interior, 2012.

ROYSTON, P. Multiple imputation of missing values. *The Stata Journal*, SAGE Publications Sage CA: Los Angeles, CA, v. 4, n. 3, p. 227–241, 2004.

RUBIN, D.; WILEY, J. New york chichester brisbane toronto singapore s. *Multiple imputation for nonresponse in surveys*, 1987.

RUBIN, D. B. Multiple imputation after 18+ years. *Journal of the American statistical Association*, Taylor & Francis, v. 91, n. 434, p. 473–489, 1996.

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.

SCHMIDT, C. *Modelo de regressão de Poisson aplicado à área da saúde*. [S.l.]: Ijúi, 2003.

STEKHOVEN, D. J.; BÜHLMANN, F. E. Missforest - non-parametric missing value imputation using random forests. *Journal of Computational and Graphical Statistics*, v. 31, n. 2, p. 413–433, 2012. Disponível em: ⟨http://www.jcgs.uni-muenster.de/doi/10.1007/jcgs.31.02⟩.

VINHA, L. G. d. A. Estudos longitudinais e tratamento de dados ausentes em avaliações educacionais. Tese (Doutorado em Psicologia), Instituto de Psicologia, Universidade de Brasília, 2016.

WHITE, I. R.; ROYSTON, P.; WOOD, A. M. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, Wiley Online Library, v. 30, n. 4, p. 377–399, 2011.