UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ITALO ALVES CARNEIRO

Regressão linear robusta auto-organizada aplicada a dados intervalares

Recife

2024

ITALO ALVES CARNEIRO

Regressão linear robusta auto-organizada aplicada a dados intervalares

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de Mestre Profissional em Ciência da Computação.

Concentration Area: Inteligência Computacional.

Supervisor: Renata Maria Cardoso Rodrigues de Souza

Co-supervisor: Leandro Carlos de Souza

Recife

2024

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Carneiro, Italo Alves.
Regressão linear robusta auto-organizada aplicada a dados intervalares / Italo Alves Carneiro. - Recife, 2024.
51f.: il.

Dissertação (Mestrado), Universidade Federal de Pernambuco, Centro de Informatica, Programa de Pós-Graduação em Ciência da Computação, 2024.
Orientação: Renata Maria Cardoso Rodrigues de Souza.
Coorientação: Leandro Carlos de Souza.

1. Análise Simbólica de Dados; 2. Dados de intervalares; 3. Regressão robusta; 4. Outliers. I. Souza, Renata Maria Cardoso Rodrigues de. II. Souza, Leandro Carlos de. III. Título.

UFPE-Biblioteca Central CDD 006.31

Dedico este trabalho a Deus, minha esposa, minha família, meus amigos e aos professores que me apoiaram durante todo o tempo.

AGRADECIMENTOS

Em primeiro lugar, a Deus, que fez com que meus objetivos fossem alcançados durante todos os meus anos de estudos.

Aos meus orientadores, Dra. Renata Souza e Dr. Leandro de Souza pelos ensinamentos e orientações que me permitiram construir este trabalho.

À minha esposa Karynne Rachel Vieira Guimarães, que me deu todo o apoio para conseguir desenvolver esse trabalho.

À minha família, que apoiou durante todo o período em que me dediquei a este trabalho.

Aos meus amigos, pelas contribuições e troca de aprendizados que contribuíram de alguma forma para a realização deste trabalho.

RESUMO

Dados simbólicos são tipos de dados complexos e podem ser representados de diferentes maneiras, cada uma com suas peculiaridades e aplicações. Dados do tipo intervalo, por sua vez, podem ser usados para representar informações imprecisas, como medições, informações sensíveis, como geolocalização, ou até mesmo como uma forma de reduzir o tamanho do problema. No contexto da regressão linear de dados intervalares, podemos ter dois problemas: sensibilidade a dados discrepantes (outliers) e escolha adequada para a representação dos intervalos. Para superar essas dificuldades, este artigo propõe um modelo de regressão robusta para dados intervalares no qual a melhor representação para os dados intervalares é obtida automaticamente, otimizando um critério baseado na equação paramétrica da reta e no método dos mínimos quadrados reponderados. Conjuntos de dados sintéticos e reais são considerados para validar o desempenho do modelo proposto. Conceitos de outliers de limite inferior e superior também foram introduzidos, além da métrica MMRE:L para limites inferiores e MMRE:U para limites superiores.

Palavras-chave: Análise Simbólica de Dados; Dados intervalares; Outliers; Regressão robusta.

ABSTRACT

Symbolic data are complex data types and can be represented in different ways, each with its peculiarities and applications. Interval-type data, in turn, can be used to represent imprecise information such as measurements, sensitive information such as geolocation or even as a way to reduce the size of the problem. From the aspect of linear regression of interval data, we can have two problems: sensitivity to outlier data and suitable choice for representing intervals. In order to overcome these difficulties, this paper proposes a robust regression model for interval data in which the best representation for interval data is obtained automatically optimizing a criterion based on the parametric equation of the line and reweighted least squares method. Synthetic and real data sets are considered in order to validate the performance of the proposed model. Concepts of lower and upper limit outliers were also introduced, in addition to the MMRE:L metric for lower limits and MMRE:U for upper limits.

Keywords: Symbolic Data Analysis; Interval data; Outliers; Robust Regression.

LISTA DE FIGURAS

LISTA DE TABELAS

CONTEÚDO

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

O objetivo principal de um modelo preditivo é representar efetivamente os dados de entrada e generalizar os padrões aprendidos para aplicação em dados futuros e não vistos. A qualidade da representação dos dados influencia significativamente o desempenho de aprendizagem do modelo (NAJAFABADI et al., 2015). Além disso, o volume substancial de dados gerado na era atual exige uma redução na complexidade e dimensionalidade sem comprometer a qualidade da informação, diminuindo assim o custo computacional associado a esses modelos.

Uma abordagem em que fundimos, resumimos e agregamos grandes e complexos conjuntos de dados em elementos como intervalos, histogramas ou polígonos, visando reduzir a complexidade ou preservar a privacidade das informações, pode ser de grande utilidade. Dessa forma, a análise de dados simbólicos (SDA) (DIDAY; NOIRHOMME-FRAITURE, 2008)(BOCK; DIDAY, 1999) lida com abordagens para dados estruturados no formato dos elementos mencionados acima. Por exemplo, dados simbólicos do tipo intervalo são um dos mais populares neste campo de estudo. Dados intervalares podem ser usados em situações de dados imprecisos, como medições, agregando informações incompletas, diminuindo o volume de dados e preservando a privacidade, como em geolocalização.

Um amplo campo de uso de dados está na tomada de decisões por meio de previsões. Uma abordagem popular é a regressão, que estabelece uma relação entre variáveis independentes e preditivas para prever cenários. Na aplicação da regressão, surge um desafio comum com a presença de um ou mais outliers nos dados. Esses outliers carregam informações valiosas sobre o ajuste do modelo e a qualidade dos dados, servindo como indicadores de fenômenos atípicos. A presença desses pontos de dados pode ter um impacto substancial na análise estatística, especialmente em modelos de regressão que dependem de estimadores de mínimos quadrados (FAGUNDES; SOUZA; CYSNEIROS, 2013).

Ao lidar com dados contaminados por outliers, a regressão robusta surge como uma alternativa à regressão de mínimos quadrados ordinários (OLS). Essa abordagem envolve a modificação da função objetivo do OLS para mitigar o impacto das observações de outliers, melhorando assim a precisão das estimativas dos parâmetros e/ou capacidades preditivas. Vários modelos foram propostos para a regressão robusta aplicada a dados intervalares (FAGUNDES; SOUZA; CYSNEIROS, 2013) (ZHAO; WANG; WANG, 2023) (NETO; CARVALHO, 2018).

Esses métodos de regressão robusta assumem a representação por ponto médio e amplitude para dados intervalares e dois modelos de regressão são executados. No entanto, essa representação pode não ser a melhor. Nesse contexto, (SOUZA et al., 2017) propuseram uma abordagem que encontra a melhor representação para as variáveis intervalares simbólicas em modelos de regressão, minimizando o critério baseado no método de mínimos quadrados.

1.2 OBJETIVO

Este trabalho tem como objetivo propor uma regressão linear robusta para dados simbólicos do tipo intervalo, na qual a melhor representação é obtida através da equação paramétrica da reta, otimizando o critério baseado no método dos mínimos quadrados reponderados iterativos (IRLS). Assim, é introduzido um método de regressão linear robusta auto-organizada para dados intervalares. Os dados intervalares são ponderados iterativamente com base em quão bem comportados eles são através de sua representação. Dois problemas independentes de minimização são resolvidos para encontrar os melhores pontos que representam os intervalos. Nesse contexto, ao encontrar a melhor representação para dados intervalares, uma nova definição de outlier é introduzida com base nos limites da variável resposta intervalar. Uma avaliação experimental com dados intervalares sintéticos e reais é considerada para mostrar a utilidade do modelo apresentado, baseada em comparações com a métrica da magnitude média do erro relativo.

Mais especificamente, este trabalho visa:

- Propor um novo método de Regressão Linear Robusta para dados de intervalo usando a representação do intervalo pela equação paramétrica da reta;
- Propor uma definição para outliers de limite inferior e superior, bem como, analisar os dados reais de acordo com a definição proposta;
- Propor métricas de avaliação para limite inferior e superior, $MMRE : L$ e $MMRE : U$;
- Avaliar o método proposto com experimentos em dados sintéticos e reais.

1.3 METODOLOGIA

A metodologia aplicada para realizar este trabalho é formada pelos seguintes pontos:

- Revisão da literatura sobre regressão linear robusta para dados intervalares, dados outliers e análise de dados simbólicos;
- Desenvolver o método de Regressão linear robusta auto-organizada;
- Definir outlier de limite inferior e superior;
- Desenvolver experimentos com dados sintéticos sobre 6 cenários deferentes;
- Desenvolver experimentos com dados reais;
- Analisar os candidatos a outliers sobre a definição proposta.

1.4 RESULTADOS ESPERADOS

Desenvolver um novo método de regressão linear robusta para dados de intervalo, utilizando a abordagem da representação do intervalo pela equação paramétrica da reta. O método proposto deve ter um desempenho satisfatório para a métrica de $MMRE$, $MMRE : L$ e $MMRE : U$, em outras palavras, deve obter resultados melhores ou estatisticamente similar aos modelos propostos anteriormente.

1.5 ESTRUTURA DO DOCUMENTO

Esta dissertação é composta por este capítulo introdutório e mais quatro capítulos. No Capítulo 2, são apresentadas as bases teóricas da regressão linear robusta para dados intervalares, dados outliers e análise de dados simbólicos; o Capítulo 3 apresenta a metodologia, que explicará o método proposto, bem como, definição de outliers de limite inferior e superior. No Capítulo 4, são definidas as métricas usadas, além de realizada experimentos e análises dos resultados obtidos e, assim, avaliar o modelo proposto e a comparação com os métodos existentes na literatura. Finalmente, no Capítulo 5, são apresentadas as conclusões e considerações finais, juntamente com sugestões para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 ANÁLISE DE DADOS SIMBÓLICOS

A Análise de Dados Simbólicos (SDA) é uma área emergente no campo da estatística e da análise de dados, que busca compreender e modelar tabelas de dados complexas, onde cada célula expressa a variabilidade de cada unidade observada. Essa variabilidade dos dados surge da agregação de observações de tabelas de dados clássicas, nas quais cada indivíduo, denominado unidade de primeiro nível, é descrito por microdados de variáveis clássicas (BRITO; DIAS, 2022). Conforme mencionado por (BRITO; DIAS, 2022), essas agregações podem ser:

- **Temporal**, se o tempo for o critério de agregação e os registros estiverem agrupados em uma unidade de tempo, por exemplo, um dia. Nesse caso, as entidades em análise são as unidades originais de primeiro nível, agora caracterizadas por conjuntos de valores provenientes dos registros recolhidos ao longo de uma unidade de tempo.
- **Contemporâneo**, se os registros forem coletados no mesmo instante temporal ou se o instante temporal não for relevante. Nas situações em que a agregação é contemporânea, as unidades de nível superior das entidades são classes de indivíduos (conjuntos de unidades de primeiro nível) agrupadas de acordo com características específicas. As variáveis que descrevem as unidades de nível superior e as respectivas unidades de primeiro nível são as mesmas; entretanto, os "valores"que as variáveis assumem para cada unidade de nível superior são agora conjuntos de valores ou distribuições obtidos das respectivas unidades de primeiro nível.

Como resultado dessa agregação, obtemos novos tipos de dados, apelidados de “simbólicos”, pois não podem ser reduzidos a números sem sacrificar muita informação. Existem diversos tipo de dados simbólicos estudados na literatura, tais como: intervalo, histograma, poligonal, entre outros. Uma aplicação notável da SDA é a detecção de fraudes, desperdícios e abusos em dados de sinistros de seguros de saúde. Estudos como o de (REYNOLDS, 2020) demonstram que técnicas de SDA podem ser eficazes na identificação de comportamentos anômalos em dados de seguros, utilizando métodos que analisam os dados em um nível conceitual superior, em vez do nível de linha tradicional

2.1.1 Dados de Intervalo

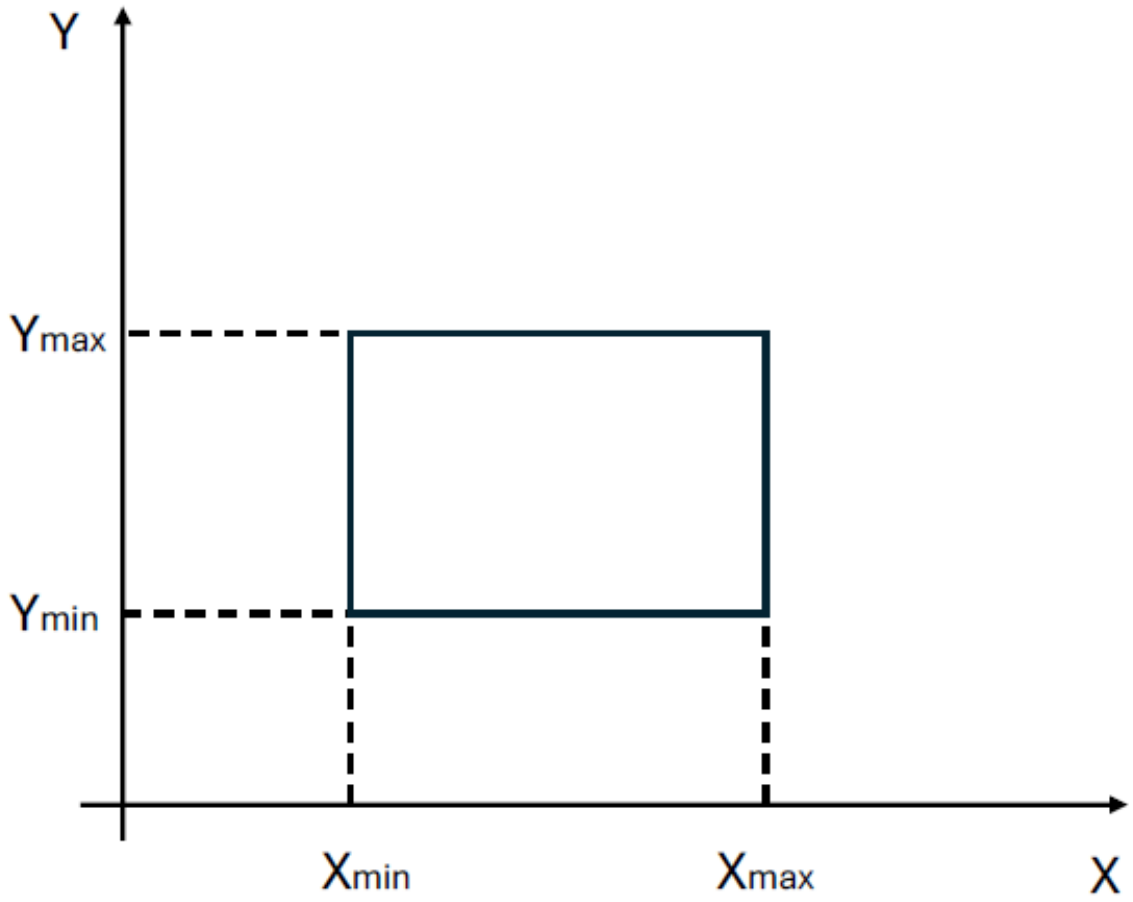
Os dados de intervalo são uma forma de agregação dos dados simbólicos em que cada observação é representada por um intervalo de valores possíveis. Esse intervalo pode ser composto pelo seu limite inferior e superior, exemplo $Y = [10, 20]$, em que 10 é seu limite inferior e 20 o superior. Existem outras representações do intervalo como centro e amplitude, equação da reta e entre outro. Para ilustrar a geração desses dados vamos trazer dois exemplos do uso de dados intervalares:

- Na criação de uma base de dados com preços de modelos de Smartphones, para cada modelo foram capturados seus limites inferiores e superiores, para cada mês do ano, assim, uma observação de um certo modelo em um certo mês pode ser expressa por $P_1 = [1.230,00; 1.450,00]$. Uma outra observação para o mesmo modelo em um outro mês de muitas promoções, teremos possíveis dados que fogem do comum, outliers, $P_2 = [920,00; 1.470,00]$. Note que em ambos há a variabilidade do produto no mês porém ao comparar com o outro período, há uma diferença no valor do limite inferior devido às promoções naquele período.

- Na medição em uma construção, pode haver divergências de medidas devido ao equipamento usado na medição. Cada equipamento possui seu grau de incerteza e por esse motivo, a medição pode ser dada por um intervalo. Se a medição $M_1 = 78cm$ feito por um equipamento que possui uma incerteza de $2cm$, logo podemos dizer que $M_1 = [76cm; 80cm]$. Dessa forma, é possível capturar a incerteza na medição usando essa abordagem de dados simbólicos.

A figura 1 ilustra graficamente um intervalo no R^2 das variáveis X e Y. A região retangular delimitada por X_{min}, X_{max}, Y_{min} e Y_{max} é o intervalo da observação.

Figura 1 – Exemplo gráfico de um intervalo no R^2.



Fonte: O autor (2024)

A tabela 1 ilustra um conjunto de dados com variáveis intervalar de estatísticas de desempenho de jogadores de basquete (WANG; GUAN; WU, 2012). As variáveis são: o número de pontos por minuto, o número de auxílios para pontuação por minuto, bem como o tempo jogado. A agregação de 96 jogadores desse conjunto de dados é feita pela sua faixa etária resultando em 13 intervalos.

Tabela 1 – Conjunto de dados Basquete

índice	idade(em anos)	pontos por minuto	auxílios por minuto	tempo jogado
1	Menor ou igual a 23	[0.2683,0.5437]	[0.0528,0.2244]	[11.81,36.55]
2	24	[0.2381,0.5668]	[0.1010,0.2282]	[10.08,33.88]
3	25	[0.3004,0.5059]	[0.0805,0.2495]	[12.63,35.22]
4	26	[0.2719,0.5769]	[0.0747,0.2383]	[17.41,38.80]
5	27	[0.2578,0.5523]	[0.0728,0.2681]	[17.46,39.53]
6	28	[0.2894,0.5885]	[0.0888,0.2771]	[18.49,38.40]
7	29	[0.4007,0.6244]	[0.1227,0.2521]	[27.87,38.43]
8	30	[0.3498,0.8291]	[0.0896,0.2130]	[12.24,40.71]
9	31	[0.2185,0.5835]	[0.0550,0.3437]	[12.12,34.91]
10	32	[0.1593,0.6318]	[0.0494,0.2327]	[13.37,36.52]
11	33	[0.2406,0.4035]	[0.1317,0.1528]	[16.36,17.46]
12	34	[0.3890,0.6318]	[0.0898,0.1236]	[13.37,28.81]
13	Maior ou igual a 35	[0.2471,0.2989]	[0.1668,0.2127]	[14.38,14.57]

2.2 REGRESSÃO LINEAR INTERVALAR

A regressão linear intervalar é uma extensão dos métodos tradicionais de regressão que lida com variáveis representadas por intervalos de valores, ao invés de valores pontuais. Essa abordagem é particularmente útil quando há incerteza ou variabilidade intrínseca nos dados, permitindo uma modelagem mais precisa e informativa.

Os modelos de regressão linear intervalar adaptam os conceitos tradicionais de mínimos quadrados e verossimilhança para acomodar intervalos, fornecendo ferramentas estatísticas robustas que capturam a variabilidade interna dos dados (FREITAS et al., 2024). Isso resulta em previsões mais realistas e uma melhor compreensão das relações entre as variáveis estudadas, considerando tanto a centralidade quanto a dispersão dos dados representados por intervalos (FREITAS et al., 2024).

Podemos encontrar nove métodos básicos diferentes de regressão linear intervalar na literatura. Muitos desses métodos, especialmente os mais antigos, utilizam análise de regressão de valor único aplicada a diferentes parâmetros do intervalo; os mais recentes, porém, como (SINOVA et al., 2012) ou (SOUZA et al., 2017) tentam se aprofundar e utilizar técnicas como a aritmética de intervalos ou sua parametrização.

Este capitulo apresenta uma revisão dos métodos de regressão linear presentes na literatura. Esses métodos são: Método do Centro (CM) (BILLARD; DIDAY, 2000), Método do Mínimo e

Máximo (MinMax) (BILLARD; DIDAY, 2002), Método do Centro e Amplitude (CRM) (NETO; CARVALHO, 2008), Método do Centro e Amplitude com Restrições (NETO; CARVALHO, 2010), Método da Informação Completa (MIC) (WANG; GUAN; WU, 2012), Método da Regressão Linear Simples baseada em aritmética (ABSLR) (SINOVA et al., 2012), Método da Regressão Linear Baseada na Técnica Lasso (GIORDANI, 2015), Método Conjunto de Centro e Amplitude Restrita (CCRJM) (HAO; GUO, 2017) e Método Parametrizado para Regressão Linear de Dados Intervalados (SOUZA et al., 2017).

2.2.1 Método do Centro (CM)

O método do Centro (CM) foi introduzido por (BILLARD; DIDAY, 2000) e utiliza os centros dos intervalos para construir um modelo de regressão linear clássico. Os coeficientes do modelo construído e os limites dos regressores são usados na predição dos limites da resposta. Os limites, inferior e superior, dos regressores determinam os limites da resposta. Este modelo foi o primeiro a ser introduzido na literatura e por isso há perda de informações ao reduzir o intervalo ao seu centro, não trazendo a informação da amplitude, por exemplo.

2.2.2 Método do mínimo e máximo (MinMax)

O Método do mínimo e máximo (MinMax) foi introduzido por (BILLARD; DIDAY, 2002) e propõe independência entre os limites inferiores e superiores, gerando dois modelos diferentes. Dessa forma, há um modelo para cada limite, composto pelo seus limites da variável resposta e aos limites da variável regressores. A equação 2.1 ilustra a modelagem dos limites:

$$\begin{cases} \underline{Y}_i = \beta_0^{inf} + \sum_{j=1}^{p} \beta_j^{inf} \underline{x}_{ji} + \varepsilon_i^{inf} \\ \overline{Y}_i = \beta_0^{sup} + \sum_{j=1}^{p} \beta_j^{sup} \overline{x}_{ji} + \varepsilon_i^{sup} \end{cases} \tag{2.1}$$

em que β representa os coeficientes da regressão, ε o seu erro e a linha inferior e superior à variável, representando o limite inferior e superior, respectivamente. A equação 2.2 ilustra a estimativa dos limites.

$$\begin{cases} \underline{\hat{Y}} = \underline{\hat{X}} \hat{\beta}^{inf} \\ \overline{\hat{Y}} = \overline{\hat{X}} \hat{\beta}^{sup} \end{cases} \tag{2.2}$$

2.2.3 Método do Centro e Amplitude (CRM)

Introduzido por (NETO; CARVALHO, 2008), o método do centro e amplitude parte da mesma premissa do método do MinMax em que temos dois modelos independentes, porém com outra representação do intervalo. Assim, teremos um modelo para o centro e outro para a amplitude do intervalo. A equação 2.3 ilustra a modelagem da regressão:

$$\begin{cases} Y_i^c = \beta_0^c + \sum_{j=1}^{p} \beta_j^c x_{ji}^c + \varepsilon_i^c \\ Y_i^a = \beta_0^a + \sum_{j=1}^{p} \beta_j^a x_{ji}^a + \varepsilon_i^a \end{cases} \tag{2.3}$$

em que β representa os coeficientes da regressão, ε o seu erro e o índices c e a representam o centro e a amplitude respectivamente. A equação 2.4 ilustra a estimativa dos limites.

$$\begin{cases} \hat{Y}^c = \hat{X}^c \hat{\beta}^c \\ \hat{Y}^a = \hat{X}^a \hat{\beta}^a \end{cases} \tag{2.4}$$

2.2.4 Método do Centro e Amplitude com restrições (CCRM)

O Método do Centro e Amplitude com Restrições que foi introduzido por (NETO; CARVALHO, 2010) propõe uma melhoria ao CRM. Esse método, assim como o CRM, também possui dois modelos independentes para o centro e amplitude, porém os coeficientes para a modelagem das amplitudes têm restrições, que os forçam a ser positivos, pois matematicamente não faz sentido uma amplitude negativa.

A restrição sugerida por (NETO; CARVALHO, 2010) para o seu método não pode ser estimada diretamente, assim, foi sugerido o uso de algoritmo iterativo, desenvolvido por (LAWSON; HANSON, 1974).

2.2.5 Método da informação completa (MIC)

(WANG; GUAN; WU, 2012) propôs uma abordagem em que a modelagem da regressão linear para dados de intervalos é feita usando toda a informação do intervalo. Para o método da informação completa (MIC), todos os pontos internos dos intervalos são usados para determinar as estimativas para os coeficientes da regressão. Assim, as variáveis regressoras geram hiper-

cubos no espaço R^P e o modelo da regressão é construído baseando-se neles. Além disso, foi adotada a combinação linear de Moore (MOORE, 1966) para garantir a coerência matemática dos limites preditos.

2.2.6 Método da Regressão linear simples baseada em aritmética (ABSLR)

O Método da Regressão Linear Simples Baseada em Aritmética (ABSLR) proposta por (SINOVA et al., 2012) é uma abordagem que utiliza operações aritméticas específicas para lidar com a natureza intervalar dos dados. No ABSLR, os dados intervalares são tratados diretamente sem transformá-los em pontos únicos, preservando assim a natureza intervalar das observações. Para um dado intervalar $[\underline{X}, \overline{X}]$, as operações aritméticas específicas são aplicadas para estimar os coeficientes de regressão.

2.2.7 Método da Regressão linear baseada na técnica Lasso

O método proposto por (GIORDANI, 2015) é baseado na Técnica Lasso (*Least Absolute Shrinkage and Selection Operator*) que é uma abordagem de regularização que realiza seleção de variáveis e encolhimento de coeficientes em modelos de regressão. Quando aplicada a dados intervalares, a Técnica Lasso pode ser adaptada para lidar com a incerteza dos intervalos enquanto ainda controla a complexidade do modelo. Como a Técnica aborda um problema de otimização envolvendo a minimização restrita de uma função objetivo, este método pode assemelhar-se à abordagem MCR. O método tentará encontrar um conjunto compartilhado de coeficientes de regressão para o ponto médio e o raio. Isto será conseguido adicionando coeficientes de regressão específicos para os raios, a fim de lidar adequadamente com todas as situações em que a inclinação difere da propagação da imprecisão (AL-ASADI, 2022).

2.2.8 Método Conjunto de centro e amplitude restrita (CCRJM)

Proposto por (HAO; GUO, 2017), o Método Conjunto de Centro e Amplitude Restrita (CCRJM) é um avanço em relação ao método CCRM, procura um coeficiente para o centro e o raio como o CCRM faz, mas o faz em conjunto, utilizando os valores dos raios para o modelo de centro e os centros para o modelo de raio. Esse método também tem como objetivo minimizar o impacto de outliers (AL-ASADI, 2022).

2.2.9 Método Parametrizado para Regressão Linear de Dados Intervalados

O Método Parametrizado para Regressão Linear de Dados Intervalados proposto por (SOUZA et al., 2017), baseia-se em uma representação do intervalo pela parametrização da reta e, como o método de MinMax, propõe coeficientes para os limites inferiores e superiores usando de forma conjunta a informação de ambos.

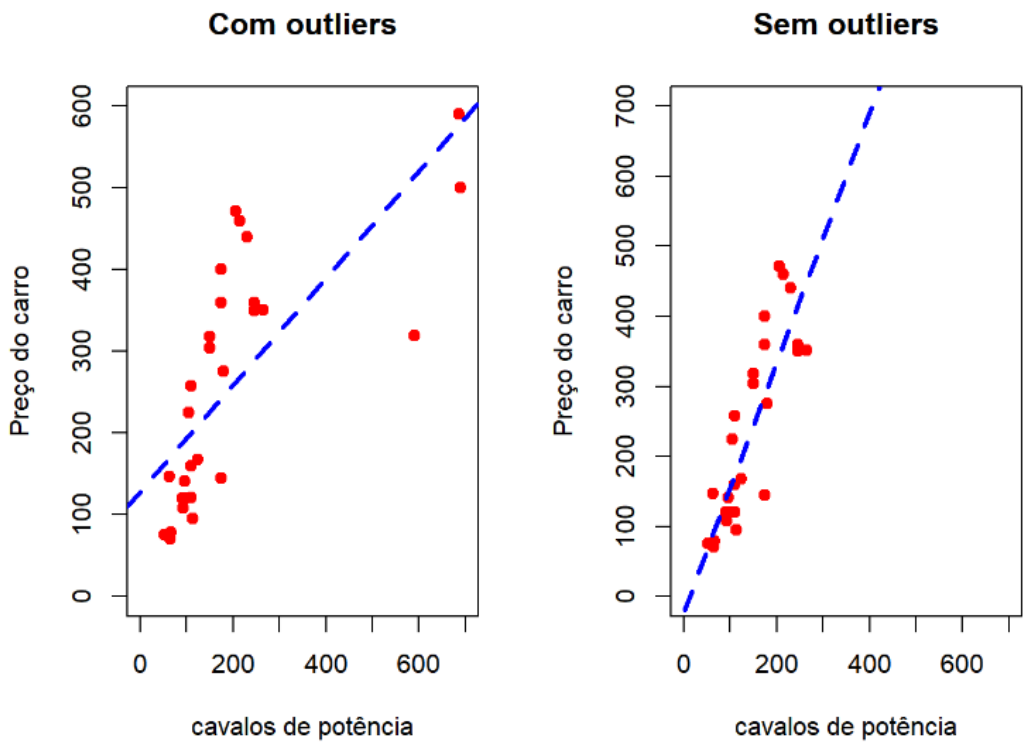
Esse método dá um passo à frente porque não apenas leva em conta o fato de que os dados com valor de intervalo têm múltiplos componentes, como os métodos anteriores, mas também reconhece o relacionamento entre esses componentes (AL-ASADI, 2022).

O método proposto neste trabalho faz uma abordagem robusta a dados outliers do método parametrizado para regressão linear de dados intervalados.

2.3 REGRESSÃO LINEAR ROBUSTA PARA DADOS DE INTERVALO

Os dados discrepantes (outliers) podem distorcer as previsões da regressão linear. A figura 2 ilustra graficamente o impacto na reta regressora em casos com outliers e sem. Nessa figura foi introduzido manualmente outliers no conjunto de dados de carros e feito uma regressão linear simples. Dessa forma, podemos notar que a reta regressora na figura com outliers fica com uma inclinação menor devido aos três pontos de outliers introduzidos no conjunto de dados.

Figura 2 – Impacto do outlier.



Fonte: <https://livro.metodosquantitativos.com/docs/outliers.html>

Como uma alternativa à regressão de mínimos quadrados ordinários (OLS), a técnica de regressão linear robusta é importante para analisar dados contaminados com outliers. Os métodos de regressão robusta para dados de intervalo visam penalizar a contribuição dos dados de outliers na regressão usando técnicas como: modificar na função objetiva da OLS, uso do mínimos quadrados reponderados (FAGUNDES; SOUZA; CYSNEIROS, 2013), uso de Kernel (NETO; CARVALHO, 2018) (CARVALHO; NETO; ROSENDO, 2021), Regressão Quantílica (SOARES; FAGUNDES, 2018), técnicas baseadas em ponto central e transformação logarítmica para amplitude (ZHAO; WANG; WANG, 2023), técnicas de dados difusos (FERRARO; GIORDANI, 2013).

2.3.1 Regressão Robusta Intervalar (IRR)

O método IRR proposto por (FAGUNDES; SOUZA; CYSNEIROS, 2013) considera duas regressões lineares robustas independentes para o centro e a amplitude dos intervalos. Além disso, a predição dos limites inferiores e superiores dos novos intervalos é baseada nas estimativas do centro e da amplitude desses intervalos.

A técnica para penalizar os outliers do conjunto de dados é um procedimento iterativo, o

mínimos quadrados reponderados usando a função de critério *Tukey's biweight*.

2.3.2 Regressão robusta do kernel do tipo exponencial para variáveis com valor de intervalo (iETKRR)

O modelo de regressão robusta do tipo exponencial para variáveis intervalares (iETKRR), proposto por (NETO; CARVALHO, 2018), utiliza a técnica de Kernel para mapear os dados intervalares em um espaço de características de alta dimensão, onde a relação entre as variáveis pode ser modelada de forma não linear. A função de Kernel do tipo exponencial é escolhida por sua capacidade de lidar com variabilidade e outliers nos dados. Além disso, o iETKRR pode ser utilizado com a representação do intervalo de centro e amplitude ou limite inferior e superior.

2.3.3 Regressão robusta conjunta de intervalo (iJRR)

O método proposto por (CARVALHO; NETO; ROSENDO, 2021), é capaz de levar em consideração as inter-relações entre os centros e os raios (ou as inter-relações entre os limites inferior e superior). Segundo o autor do método, a abordagem proposta é possivelmente o primeiro método de regressão robusto que leva em conta a informação completa do intervalo nos modelos de regressão ajustados.

O método iJRR baseia-se numa função objetivo adequada de dois termos com o objetivo de levar em consideração conjuntamente a informação fornecida quer pelo centro e pelo raio, quer pelos limites inferior e superior dos intervalos. Consequentemente, o método iJRR ajusta dois modelos de regressão, seja no centro e no intervalo ou nos limites inferior e superior dos intervalos.

Além disso, é usado uma função de kernel Gaussiana com o mesmo objetivo do método iETKRR e também pode ser utilizado com a representação do intervalo de centro e amplitude ou limite inferior e superior.

2.3.4 Regressão quantílica intervalar (IQR)

O método IQR, como proposto por (SOARES; FAGUNDES, 2018), representa uma abordagem inovadora na regressão quantílica. Este método proporciona estimativas fundamentadas

na mediana condicional e em uma ampla gama de outras funções quantílicas, mediante a minimização dos erros absolutos ponderados. A complexidade na estimativa dos coeficientes do IQR é reconhecidamente maior em comparação a outros modelos de regressão, dado que não existe uma forma exata para essa estimativa. Os algoritmos mais frequentemente utilizados para a tarefa de estimativa de coeficientes são o simplex e o ponto interior, ambos fundamentados em Programação Linear e pertencentes à classe de algoritmos determinísticos.

Ademais, no contexto do IQR, são empregados algoritmos probabilísticos baseados em Swarms com o intuito de simplificar a estimativa dos coeficientes de regressão quantílica e melhorar o desempenho do modelo no ajuste de dados simbólicos intervalares. Para esse método, a representação de centro e amplitude é também utilizada.

2.3.5 Regressão Robusta intervalar baseada em ponto central e logaritmo da amplitude (LN-IRR)

O método proposto por (ZHAO; WANG; WANG, 2023), assim como o IRR, considera duas regressões lineares robustas independentes, porém uma para o centro e outra o logaritmo da amplitude dos intervalos. De forma similar, é usado um algoritmo de mínimos quadrados reponderados iterativamente usando a função de critério de Huber, aplicada no procedimento de estimativa para tolerar a existência de outliers na variável de intervalo dependente.

O LN-IRR também visa solucionar problemas de valores não negativos para dados de amplitude, que corresponde a uma inconsistência matemática.

2.3.6 Regressão Robusta Intervalar para dados difusos(Fuzzy-RR)

A lógica difusa é utilizada para modelar incertezas e imprecisões nos dados. Em vez de trabalhar com valores exatos, a lógica difusa permite a manipulação de graus de pertinência, que representam o quanto um dado pertence a uma determinada categoria.

A abordagem Fuzzy-RR, como proposta por (FERRARO; GIORDANI, 2013), combina as técnicas da lógica difusa e regressão robusta para melhorar a precisão e robustez do modelo de regressão. Primeiro, os dados são processados usando a lógica difusa para suavizar as incertezas e imprecisões. Em seguida, uma técnica de regressão robusta é aplicada aos dados difusos para obter o modelo final. Os resultados do modelo de regressão difusa são convertidos de volta para valores numéricos através do processo de inverso.

3 REGRESSÃO LINEAR ROBUSTA AUTO-ORGANIZADA APLICADA A DADOS INTERVALARES

Este capítulo descreve o método de parametrização do intervalo pela equação do segmento de reta, que se baseia no método dos mínimos quadrados para estimação dos coeficientes da regressão, sem considerar um comportamento probabilístico para os erros, proposto por (SOUZA et al., 2017). Além disso, será demonstrado o método proposto de Regressão linear robusta auto-organizada aplicada a dados intervalares ou *Self-Organized Interval Robust Regression* (SO-IRR), nesse método de regressão linear, são usados os limites inferiores e superiores do intervalo e aplicam-se penalidades aos potenciais outliers para a estimação dos coeficientes de regressão. Como demonstrado em (SOUZA et al., 2017), com uso da equação paramétrica da reta, os modelos determinam, automaticamente, os pontos nos regressores que oferecem o melhor ajuste na regressão, assim surge o nome "Self-Organized". Além disso, serão introduzidos os conceitos de outliers de limite inferior e superior.

Este capítulo foi dividido em 4 seções: Parametrização intervalar, Estimativa dos coeficientes, Regra de predição para dados intervalares e Outliers para dados simbólicos intervalares.

3.1 PARAMETRIZAÇÃO INTERVALAR

Esta seção apresenta o método Self-Organized Interval Robust (SO-IRR). Conforme proposto em (SOUZA et al., 2017), um intervalo tem uma representação geométrica como um segmento de linha. Dessa forma, se definirmos um intervalo como $\gamma = [\underline{\gamma}, \overline{\gamma}]$, com $\underline{\gamma} \leq \overline{\gamma}$, qualquer ponto q pertencente ao segmento de linha com limites no intervalo γ, pode ser determinado pela equação parametrizada da linha, apresentada na equação (3.1) com $0 \leq \lambda \leq 1$ (LEITHOLD, 1972)(MCCREA, 2012).

$$q(\lambda) = \underline{\gamma}(1 - \lambda) + \overline{\gamma}\lambda \tag{3.1}$$

Considere um conjunto de dados de tamanho n. Seja $Y = \{y_1, \ldots, y_n\}$ uma resposta intervalar com n intervalos $\{y_1 = [\underline{y}_1, \overline{y}_1], \ldots, y_n = [\underline{y}_n, \overline{y}_n]\}$ e p variáveis regressoras intervalares $\{X_1, \ldots, X_p\}$ com n intervalos $X_j = \{x_{1j}, \ldots, x_{nj}\}$ sendo $x_{ij} = [\underline{x}_{ij}, \overline{x}_{ij}]$. Considere um valor λ_j para cada variável intervalar independente X_j $(j = 1, \ldots, p)$. Assim, os pontos parametrizados q_{ij} x_{ij} $(i = 1, \ldots, n; j = 1, \ldots, p)$ definidos conforme dado em (SOUZA et al., 2017) são:

$$q_{ij} = \underline{x}_{ij}(1 - \lambda_j) + \overline{x}_{ij}\lambda_j \tag{3.2}$$

Para modelar os limites inferior e superior da variável dependente intervalar Y com base nesses pontos parametrizados, temos:

$$\begin{cases} \underline{y}_i = \beta_0^L + \sum_{j=1}^{p} \beta_j^L q_{ij} + \varepsilon_i^L \\ \overline{y}_i = \beta_0^U + \sum_{j=1}^{p} \beta_j^U q_{ij} + \varepsilon_i^U \end{cases} \tag{3.3}$$

onde $\boldsymbol{\beta}^L = (\beta_0^L, \ldots, \beta_p^L)^T$ and $\boldsymbol{\beta}^U = (\beta_0^U, \ldots, \beta_p^U)^T$ são coeficientes desconhecidos do modelo e ε_i^L e ε_i^U são erros $(i = 1 \ldots, n)$. Substituindo a Equação (3.2) na Equação (3.3), temos:

$$\begin{cases} \underline{y}_i = \beta_0^L + \sum_{j=1}^{p} \beta_j^L (\underline{x}_{ij}(1 - \lambda_j) + \overline{x}_{ij}\lambda_j) + \varepsilon_i^L \\ \overline{y}_i = \beta_0^U + \sum_{j=1}^{p} \beta_j^U (\underline{x}_{ij}(1 - \lambda_j) + \overline{x}_{ij}\lambda_j) + \varepsilon_i^U \end{cases} \tag{3.4}$$

Seja α_j e ω_j dois coeficientes dados por

$$\begin{cases} \alpha_j &= \beta_j(1 - \lambda_j) \\ \omega_j &= \beta_j\lambda_j. \end{cases} \tag{3.5}$$

Podemos reescrever a Equação (3.3) como

$$\begin{cases} \underline{y}_i = \beta_0^L + \sum_{j=1}^{p} (\alpha_j^L \underline{x}_{ij} + \omega_j^L \overline{x}_{ij}) + \varepsilon_i^L \\ \overline{y}_i = \beta_0^U + \sum_{j=1}^{p} (\alpha_j^U \underline{x}_{ij} + \omega_j^U \overline{x}_{ij}) + \varepsilon_i^U \end{cases} \tag{3.6}$$

Seja $\mathbf{X}$ uma matriz

$$\mathbf{X} = \begin{pmatrix} 1 & \underline{x}_{11} & \overline{x}_{11} & \ldots & \underline{x}_{p1} & \overline{x}_{p1} \\ 1 & \underline{x}_{12} & \overline{x}_{12} & \ldots & \underline{x}_{p2} & \overline{x}_{p2} \\ \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\ 1 & \underline{x}_{1n} & \overline{x}_{1n} & \ldots & \underline{x}_{pn} & \overline{x}_{pn} \end{pmatrix}$$

Os modelos de regressão parametrizados podem ser reescritos como

$$\begin{cases} \underline{\mathbf{y}} = \mathbf{X}\boldsymbol{\Delta}^L + \boldsymbol{\varepsilon}^L \\ \overline{\mathbf{y}} = \mathbf{X}\boldsymbol{\Delta}^U + \boldsymbol{\varepsilon}^U \end{cases} \tag{3.7}$$

onde $\boldsymbol{\Delta}^L$ e $\boldsymbol{\Delta}^U$ são escritos por

$$\boldsymbol{\Delta}^L = (\beta_0^L, \alpha_1^L, \omega_1^L, \ldots, \alpha_p^L, \omega_p^L)^T$$

$$\boldsymbol{\Delta}^U = (\beta_0^U, \alpha_1^U, \omega_1^U, \ldots, \alpha_p^U, \omega_p^U)^T$$

3.2 ESTIMANDO OS COEFICIENTES

De acordo com os modelos de regressão parametrizados na Equação (3.7) os vetores $\boldsymbol{\Delta}^L$ e $\boldsymbol{\Delta}^U$ são dados reorganizando a função para $\boldsymbol{\varepsilon}^L$ e $\boldsymbol{\varepsilon}^U$ por:

$$\boldsymbol{\varepsilon}^L = \underline{\mathbf{y}} - \mathbf{X}\boldsymbol{\Delta}^L$$

e

$$\boldsymbol{\varepsilon}^U = \overline{\mathbf{y}} - \mathbf{X}\boldsymbol{\Delta}^U$$

.

No qual $\boldsymbol{\varepsilon}^L = (\varepsilon_1^L, \ldots, \varepsilon_n^L)$ e $\boldsymbol{\varepsilon}^U = (\varepsilon_1^U, \ldots, \varepsilon_n^U)$ são variáveis aleatórias independentes e são estimados minimizando uma função critério baseada em ρ para ambos os vetores de erros, que é dada por

$$\sum_{i=1}^{n} \rho\left(\frac{\varepsilon_i^L}{S}\right) + \rho\left(\frac{\varepsilon_i^U}{S}\right). \tag{3.8}$$

onde S é uma estimativa robusta de escala e ρ é uma função específica. Segundo (MONTGOMERY; PECK; VINING, 2021) uma escolha popular para S é a mediana da discrepância absoluta definida como

$$S = \frac{median|\varepsilon_i - median(\varepsilon_i)|}{0.6745} \tag{3.9}$$

a constante de ajuste 0.6745 torna S um estimador aproximadamente imparcial se n for grande e a distribuição de erro for normal.

Da equação (3.8) temos dois problemas independentes de minimização:

1. encontrar $\boldsymbol{\Delta}^L$ que minimiza

$$\sum_{i=1}^{n} \rho \left(\frac{\underline{y}_i - \mathbf{X}_i \boldsymbol{\Delta}^L}{S} \right). \tag{3.10}$$

2. encontrar $\boldsymbol{\Delta}^U$ que minimiza

$$\sum_{i=1}^{n} \rho \left(\frac{\overline{y}_i - \mathbf{X}_i \boldsymbol{\Delta}^U}{S} \right). \tag{3.11}$$

onde

$$\mathbf{X}_i = (1 \ \underline{x}_{1i} \ \overline{x}_{1i} \ \ldots \ \underline{x}_{pi} \ \overline{x}_{pi})^T$$

Além disso, (MONTGOMERY; PECK; VINING, 2021) demonstra que para minimizar a 3.9 deve igualar as primeiras derivadas parciais de ρ com relação a Δ a zero, denotada por $\psi = \rho'$. Em geral a função ψ é não linear e deve ser resolvida por métodos iterativos. Existem várias funções de critério associadas ao método de regressão robusta caracterizadas por $\psi(x)$ que controla os pesos $W(x)$. Neste trabalho, consideramos a função de bi-peso de Tukey, que atribui pequenos pesos a resíduos grandes. Esta função $\rho(x)$ e sua correspondente $\psi(x)$, bem como os pesos $W(x)$ a são dadas por, respectivamente:

$$\rho(x) = \begin{cases} \frac{c^2}{6}(1 - [1(x/c)^2]^3) & \text{se} \quad |x| \leq c \\ \frac{c^2}{6} & \text{se} \quad |x| > c \end{cases} \tag{3.12}$$

$$\psi(x) = \begin{cases} x(1 - [1(x/c)^2]^2) & \text{se} \quad |x| \leq c \\ 0 & \text{se} \quad |x| > c \end{cases} \tag{3.13}$$

$$W(x) = \begin{cases} \frac{\psi[(\mathbf{y_i} - \mathbf{X_i}\boldsymbol{\Delta})/S]}{(\mathbf{y_i} - \mathbf{X_i}\boldsymbol{\Delta})/S} & \text{se} \quad \mathbf{y_i} \neq \mathbf{X_i}\boldsymbol{\Delta} \\ 1 & \text{se} \quad \mathbf{y_i} = \mathbf{X_i}\boldsymbol{\Delta} \end{cases} \tag{3.14}$$

c controla o ponto de corte da função de perda que define o quanto uma observação pode ser considerada outlier. O valor de c pode variar e $c = 4.685$ é um valor recomendado e corresponde a cerca de 95% de cobertura para uma distribuição normal.

O método mais utilizado para estimar Δ^L and Δ^U é o Método de Scoring de Fisher (mínimos quadrados reponderados com uma variável de resposta modificada) (BEATON; TUKEY, 1974). O Algoritmo 1 mostra o processo iterativo para encontrar $\hat{\Delta}^L$ e $\hat{\Delta}^U$.

Algorithm 1 Mínimos quadrados reponderados

Input: Variável resposta $\mathbf{Y}$; variável regressora $\mathbf{X}$; Número máximo de iterações $\mathbf{MaxInter}$; $end = \mathbf{False}$;

Output: $\hat{\mathbf{\Delta}}^L$ e $\hat{\mathbf{\Delta}}^U$

1: **Inicializar** $\mathbf{W^L}$ e $\mathbf{W^U}$ como duas matrizes diagonais e $t - 0$;

2: **Calcule** $\hat{\mathbf{\Delta}}^L$, $\hat{\mathbf{\Delta}}^{\mathbf{U}}$ e S inicial pelo método dos mínimos quadrados;

3: **Enquanto** número de interações $'t' < MaxInter$ **OU** $end = False$

4: **Compute** os pesos

$$w_{it}^L = \begin{cases} \frac{\psi[(\underline{\mathbf{y}}_\mathbf{i} - \mathbf{X_i}\mathbf{\Delta}_t^L)/S]}{(\underline{\mathbf{y}}_\mathbf{i} - \mathbf{X_i}\mathbf{\Delta}_t^L)/S} & \text{se} \quad \underline{\mathbf{y}}_\mathbf{i} \neq \mathbf{X_i}\mathbf{\Delta}_t^L \\ 1 & \text{se} \quad \underline{\mathbf{y}}_\mathbf{i} = \mathbf{X_i}\mathbf{\Delta}_t^L \end{cases}$$

$$w_{it}^U = \begin{cases} \frac{\psi[(\overline{\mathbf{y}}_\mathbf{i} - \mathbf{X_i}\mathbf{\Delta}_\mathbf{t}^\mathbf{U})/S]}{(\overline{\mathbf{y}}_\mathbf{i} - \mathbf{X_i}\mathbf{\Delta}_\mathbf{t}^\mathbf{U})/S} & \text{se} \quad \overline{\mathbf{y}}_\mathbf{i} \neq \mathbf{X_i}\mathbf{\Delta}_\mathbf{t}^\mathbf{U} \\ 1 & \text{se} \quad \overline{\mathbf{y}}_\mathbf{i} = \mathbf{X_i}\mathbf{\Delta}_\mathbf{t}^\mathbf{U} \end{cases}$$

5: **Compute** os coeficientes

$$\hat{\mathbf{\Delta}}_{\mathbf{t+1}}^{\mathbf{L}} = (\mathbf{X'W}_{t+1}^L\mathbf{X})^{-1}\mathbf{X'W}_{t+1}^L\underline{\mathbf{y}}$$

$$\hat{\mathbf{\Delta}}_{\mathbf{t+1}}^{\mathbf{U}} = (\mathbf{X'W}_{t+1}^U\mathbf{X})^{-1}\mathbf{X'W}_{\mathbf{t+1}}^{\mathbf{U}}\overline{\mathbf{y}}$$

6: **Se**

$$\left\| \frac{\hat{\mathbf{\Delta}}_{t+1}^L - \hat{\mathbf{\Delta}}_t^L}{\hat{\mathbf{\Delta}}_t^L} \right\| \leq \epsilon \textbf{ and } \left\| \frac{\hat{\mathbf{\Delta}}_{t+1}^U - \hat{\mathbf{\Delta}}_t^U}{\hat{\mathbf{\Delta}}_t^U} \right\| \leq \epsilon$$

7: **Faça** $end = True$

onde $\epsilon < 0.0001$ pare, senão faça $t = t + 1$ e vá para a etapa 4.

3.3 REGRA DE PREDIÇÃO PARA DADOS INTERVALARES

Seja $\mathbf{v} = (1, \underline{v}_1, \overline{v}_1, \ldots, \underline{v}_p, \overline{v}_p)^T$ um novo vetor de variáveis independentes intervalares. Dadas $\hat{\mathbf{\Delta}}^L$ e $\hat{\mathbf{\Delta}}^U$, o intervalo de resposta previsto é:

$$\begin{cases} \overline{\hat{y}} = \mathbf{v}\hat{\mathbf{\Delta}}^U \\ \underline{\hat{y}} = \mathbf{v}\hat{\mathbf{\Delta}}^L \end{cases} \tag{3.15}$$

No qual $\overline{\hat{y}}$ e $\underline{\hat{y}}$ são os valores previstos dos limites superior e inferior, respectivamente.

3.4 DEFINIÇÃO DE OUTLIERS PARA DADOS SIMBÓLICOS INTERVALARES

Outliers potenciais podem ser definidos como uma observação com valores altos de resíduos studentizados. Resíduos, de acordo com a definição clássica de regressão linear, representam a diferença entre o valor previsto pela regressão e o valor observado (MONTGOMERY; PECK; VINING, 2021). Esses valores altos de resíduos podem ser definidos como um valor maior ou igual a um determinado limiar, denotado como τ, que em nosso estudo será definido como $\tau = 2$.

No contexto de dados intervalares, (FAGUNDES; SOUZA; CYSNEIROS, 2013) propôs a primeira definição de outlier intervalar. Um outlier intervalar refere-se a um elemento de um conjunto de dados em que seu ponto médio ou amplitude ou ponto médio e amplitude da variável de resposta está a uma distância anormal dos outros valores do ponto médio desse conjunto de dados. Com relação a essa definição, várias abordagens robustas de regressão para dados intervalares foram publicadas na literatura de SDA (FAGUNDES; SOUZA; CYSNEIROS, 2013), (NETO; CARVALHO, 2018), (ZHAO; WANG; WANG, 2023).

Neste estudo, três tipos de outliers intervalares são definidos no contexto da regressão linear para dados simbólicos. Eles são baseados em informações residuais relativas ao limite inferior, limite superior e limites inferior e superior conjuntamente, respectivamente.

3.4.1 DEFINIÇÃO DE OUTLIERS PARA LIMITE INFERIOR

Um outlier intervalar (X_i, y_i) baseado no limite inferior refere-se a um objeto i para o qual $|\underline{\Delta_i}| \geq \tau$ no limite inferior da coordenada $\underline{y}$, onde

$$\underline{\Delta_i} = \frac{\underline{y_i} - \underline{\hat{y}_i}}{\underline{\sigma_i}\sqrt{1 - \underline{h_{ii}}}} \tag{3.16}$$

onde $\underline{\sigma_i}$ é chamada de média quadrática residual e pode ser estimada por (MONTGOMERY; PECK; VINING, 2021),

$$\underline{\hat{\sigma}_i} = \frac{SS_{res}}{n - p}$$

$$SS_{res} = \sum_{i=1}^{n} (\underline{y_i} - \underline{\hat{y}_i})^2$$

o termo $(n-p)$ representa o grau de liberdade associado, onde n é o tamanho da amostra e p o número de parâmetros de regressão. e $\underline{h_{ii}}$ é geralmente chamado de elemento diagonal i-ésimo da matriz chapéu $\underline{\mathbf{H}}$ para os limites inferiores, e é a matriz de projeção que expressa os valores das observações na variável dependente, $\underline{Y}$, em termos de combinações lineares dos vetores coluna da matriz do modelo, $\underline{X}$, que contém as observações.

$$\underline{\mathbf{H}} = \underline{\mathbf{x}_i}(\underline{\mathbf{X}}^T\underline{\mathbf{X}})^{-1}\underline{\mathbf{x}_i}^T$$

onde $\underline{\mathbf{X}} = (\underline{\mathbf{x}}_1, \underline{\mathbf{x}}_2, \ldots, \underline{\mathbf{x}}_n)$.

3.4.2 DEFINIÇÃO DE OUTLIERS PARA LIMITE SUPERIOR

Semelhante ao limite inferior, um outlier intervalar (X_i, y_i) baseado no limite superior refere-se a um objeto i para o qual $|\overline{\Delta_i}| \geq \tau$ no limite superior da coordenada $\overline{y}$, onde

$$\overline{\Delta_i} = \frac{\overline{y}_i - \overline{\hat{y}}_i}{\overline{\sigma}_i\sqrt{1-\overline{h}_{ii}}} \tag{3.17}$$

onde $\overline{\sigma}_i$ é chamada de média quadrática residual e pode ser estimada por (MONTGOMERY; PECK; VINING, 2021),

$$\overline{\hat{\sigma}}_i = \frac{SS_{res}}{n-p}$$

$$SS_{res} = \sum_{i=1}^{n}(\overline{y}_i - \overline{\hat{y}_i})^2$$

e $\overline{h}_{ii}$ é geralmente chamado de elemento diagonal i-ésimo da matriz chapéu $\overline{\mathbf{H}}$ para os limites superiores, e é a matriz de projeção que expressa os valores das observações na variável dependente, $\overline{Y}$, em termos de combinações lineares dos vetores coluna da matriz do modelo, $\overline{X}$, que contém as observações.

$$\overline{\mathbf{H}} = \overline{\mathbf{x}_i}(\overline{\mathbf{X}}^T\overline{\mathbf{X}})^{-1}\overline{\mathbf{x}_i}^T$$

onde $\overline{\mathbf{X}} = (\overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2, \ldots, \overline{\mathbf{x}}_n)$.

3.4.2.1 DEFINIÇÃO DE OUTLIER INTERVALAR PARA LIMITE INFERIOR E SUPERIOR

Um outlier intervalar (X_i, y_i) baseado nos limites inferior e superior refere-se a um objeto i para o qual $|\underline{\Delta_i}| \geq \tau$ e $|\overline{\Delta_i}| \geq \tau$ com $\underline{\Delta_i}$ e $\overline{\Delta_i}$ definidos como nas Equações 3.16 e 3.17, respectivamente.

4 AVALIAÇÃO E RESULTADOS

Este capitulo tem como objetivo realizar experimentos com dados simbólicos sintéticos e reais com o intuito de avaliar o modelo de regressão linear robusta para dados intervalares proposto neste trabalho, SO-IRR.

Sua organização será da seguinte forma: inicialmente, são apresentados as métricas usadas nos experimentos, o método de avaliação para os dados sintéticos usando simulação de Monte Carlo, a construção dos conjuntos de dados sintéticos assumindo dois grupos de cenários de dados, cada um com diferentes situações de outlier e seus resultados. Para os dados reais, será feito uma apresentação dos dados usados, seus experimentos e análise de candidatos a outliers nos dados reais.

4.1 MÉTRICAS DE AVALIAÇÃO

Para avaliar o método proposto neste trabalho (SO-IRR), três métricas são adotadas com base na média da magnitude do erro relativo calculado a partir dos valores do limite inferior (MMRE:L), valores do limite superior (*MMRE:U*) e valores dos limites inferior e superior conjuntamente (*MMRE*), respectivamente. Essas métricas são dadas por

$$MMRE = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\left(\left\|\frac{\underline{y}_i - \underline{\hat{y}}_i}{\underline{y}_i}\right\| + \left\|\frac{\overline{y}_i - \overline{\hat{y}}_i}{\overline{y}_i}\right\|\right) \tag{4.1}$$

$$MMRE : L = \frac{1}{n}\sum_{i=1}^{n}\left(\left\|\frac{\underline{y}_i - \underline{\hat{y}}_i}{\underline{y}_i}\right\|\right) \tag{4.2}$$

$$MMRE : U = \frac{1}{n}\sum_{i=1}^{n}\left(\left\|\frac{\overline{y}_i - \overline{\hat{y}}_i}{\overline{y}_i}\right\|\right) \tag{4.3}$$

4.2 EXPERIMENTOS COM DADOS SINTÉTICOS

As métricas descritas na seção anterior serão estimadas com base em uma simulação de Monte Carlo com 500 repetições para cada cenário do conjunto de dados sintéticos. O intervalo de confiança baseado no método de bootstrap é obtido com um intervalo de confiança de 95% para cada métrica. O algoritmo 2 descreve o experimento de Monte Carlo.

Algorithm 2 Simulação de Monte Carlo.

Input número de repetições N; porcentagem de outliers g; tipo do cenário $s \in \{1, 2, \ldots, 6\}$
Output intervalo de confiança boostrap

1: **Enquanto** $1 < i < N$;

2: **Obtenha** O i-ésimo conjuto de dados de acordo com o cenário s no Algorítimo 3;

3: **Divida** o conjunto de dados aleatoriamente em subconjuntos de treinamento e teste.;

4: **Gerar** outliers no conjunto de treino de acordo com o cenário s no Algorítimo 4

5: **Construa** o modelo de regressão para dados intervalar;

5: **Obtenha** a predição intervalar para o conjunto de teste;

6: **Calcule** $MMRE_i$, $MMRE_i : L$ e $MMRE_i : U$.

7: **fim Enquanto**

8: **Construa** o intervalo de confiança bootsrap para $MMRE$, $MMRE : L$ e $MMRE : U$.

4.3 CENÁRIOS COM DADOS SINTÉTICOS

Consideramos sete cenários de conjuntos de dados sintéticos. Os três primeiros cenários (1 a 3) são definidos como em (FAGUNDES; SOUZA; CYSNEIROS, 2013), com dados em R^3. Os dados intervalares nesses cenários são oriundos dos valores máximos e mínimos de uma distribuição normal multivariada. Para esses cenários, os outliers são obtidos em termos do centro e amplitude dos intervalos. Nos últimos três cenários (4 a 6), projetamos dados em R^2, nos quais os limites da variável de resposta dependem de pontos internos da variável preditora, conforme descrito em (SOUZA et al., 2017). Outliers de limite inferior e/ou superior são incluídos nesses cenários. O Algoritmo 3 descreve como os conjuntos de dados sintéticos são obtidos considerando diferentes configurações.

Algorithm 3 Gerando conjunto de dados sinteticos

Input tamanho com conjunto de dados: n, cenário s
Output Conjunto de dados intervalar
Se Cenário $s =$ 1 a 3 Faça

1: **Enquanto** $0 < i < n$;

3: **Gerar** um erro ε_i da distribuição normal N(0,1);

4: **Gerar** x_{i1} e $x_{i,2}$ cada um a partir de uma distribuição uniforme U[20,40];

5: **Compute** $y_i = b_0 + x_{i1}b_1 + x_{i2}b_2 + \varepsilon_i$;

6: **FimEnquanto**

7: **Enquanto** $0 < i < n$

8: **Obtenha** uma i-ésima amostra de tamanho 50 em R^3 de acordo com uma distribuição normal multivariada com vetor médio $\mu = (x_{i1}, x_{i2}, y_i)$ e a matriz de covariância diagonal $\sum$ com $\sigma_{jj} = 9$ $(j = 1, 2, 3)$;

9: **Compute** intervalos preditores $[\underline{x}_{i1} = l_1, \overline{x}_{i1} = u_1]$, $[\underline{x}_{i2} = l_2, \overline{x}_{i2} = u_2]$ e a resposta do intervalo $[\underline{y}i = l_3, \overline{y}_i = u_3]$ onde l_j e u_j $(j = 1, 2, 3)$ são, respectivamente, os valores mínimo e máximo do i-ésima amostra em R^3 do passo 7;

10: **FimEnquanto**

11: **FimSe**

Se Cenário $s =$ 4 a 6 Faça

12: **Gerar** β_0 de uma distribuição uniforme U[-67,5,-62,5]; β_1 de uma distribuição uniforme U[-5,5]; β_2 de uma distribuição uniforme U[62.5,67.5] β_3 de uma distribuição uniforme U[-5,5];

13: **Enquanto** $0 < i < n$;

14: **Gerar** x_{ci} de acordo com uma distribuição uniforme U[-10,10] e x_{ri} seguindo uma distribuição uniforme U[0,5];

15: **Compute** $\underline{x}_i = x_{ci} - x_{ri}/2$ e $\overline{x}_i = x_{ci} + x_{ri}/2$;

16: **FimEnquanto**

17: **Enquanto** $0 < i < n$;

18: **Gerar** ε_i e γ_p da distribuição normal N(0,1) e calcule $q_i = (1 - \gamma_q)\underline{x}_i + \gamma_q\overline{x}_i$ e $\underline{y}_i = \beta_0 + \beta_1 q_i + \epsilon_i$;

19: **Gerar** ε_i da distribuição normal N(0,1) e calcule $r_i = (1-\gamma_r)\underline{x}_i+\gamma_r\overline{x}_i$ e $\overline{y}_i = \beta_2+\beta_3 r_i+\epsilon_i$;

20: **FimEnquanto**

21: **FimSe**

Algorithm 4 Gerando outliers.

Input Conjunto de dados, porcentagem de outliers: k, cenário s;
Output Conjunto de dados com $k\%$ outliers

Se cenário $s =$ 1 a 3 Faça

1: **Enquanto** $0 < i < n$, onde n é o tamanho do conjunto de dados;

2: **Compute** $y_i^c = (\underline{y}_i + \overline{y}_i)/2$, $x_{i1}^c = (\underline{x}_{i1} + \overline{x}_{i1})/2$, $x_{i2}^c = (\underline{x}_{i2} + \overline{x}_{i2})/2$;

3: **Compute** $y_i^r = \overline{y}_i - \underline{y}_i$, $x_{i1}^r = \overline{x}_{i1} - \underline{x}_{i1}$, $x_{i2}^r = \overline{x}_{i2} - \underline{x}_{i2}$;

4 **FimEnquanto**

5: **Se Cenário 1** do conjunto de dados $\{y_i^c, x_{i1}^c, x_{i2}^c\}$ $(i = 1, \ldots, n)$ em ordem decrescente, obtenha o primeiro elemento no ponto médio classificado do conjunto de dados. Calcule valores discrepantes de ponto médio por $y_i^c = y_i^c + 9S_{y^c}$ com $(i = 1, ..., k)$ onde S_{y^c} é o desvio padrão do conjunto $(y_i^c, ..., y_n^c)$.

6: **Se Cenário 2** Do conjunto de dados $\{y_i^c, x_{i1}^c, x_{i2}^c\}$ $(i = 1, \ldots, n)$ selecione k elementos aleatoriamente e calcule o outlier do intervalo por $y_i^r = y_i^r 9S_{y^c}$ com $(i = 1, \ldots, k)$;

7: **Se Cenário 3** Do conjunto de dados $\{y_i^c, x_{i1}^c, x_{i2}^c\}$ $(i = 1, \ldots, n)$ em ordem decrescente. Calcule k outliers de ponto médio por $y_i^c = y_i^c + 9S_{y^c}$ com $(i = 1, ..., k)$. Selecione aleatoriamente uma amostra de 30% de elementos do conjunto de valores discrepantes do ponto médio e calcule os valores discrepantes do intervalo por $y_i^r = y_i^r 9S_{y^c}$ com $(i = 1, \ldots, 30\%k)$.

FimSe

Se Cenario $s =$ 4 to 6 Do

1: **Se Cenário 4** Do conjunto de dados $\{\underline{y}_i, \underline{x}_{i1}, \underline{x}_{i2}\}$ $(i = 1, \ldots, n)$ em ordem decrescente, obtenha o primeiro elemento do conjunto de dados do ponto médio classificado. Calcule valores discrepantes de limite inferior por $\underline{y}_i = \underline{y}_i - 9S_{\underline{y}}$ com $(i = 1, ..., k)$ onde $S_{\underline{y}}$ é o desvio padrão do conjunto $(\underline{y}_i, \ldots, \underline{y}_n)$.

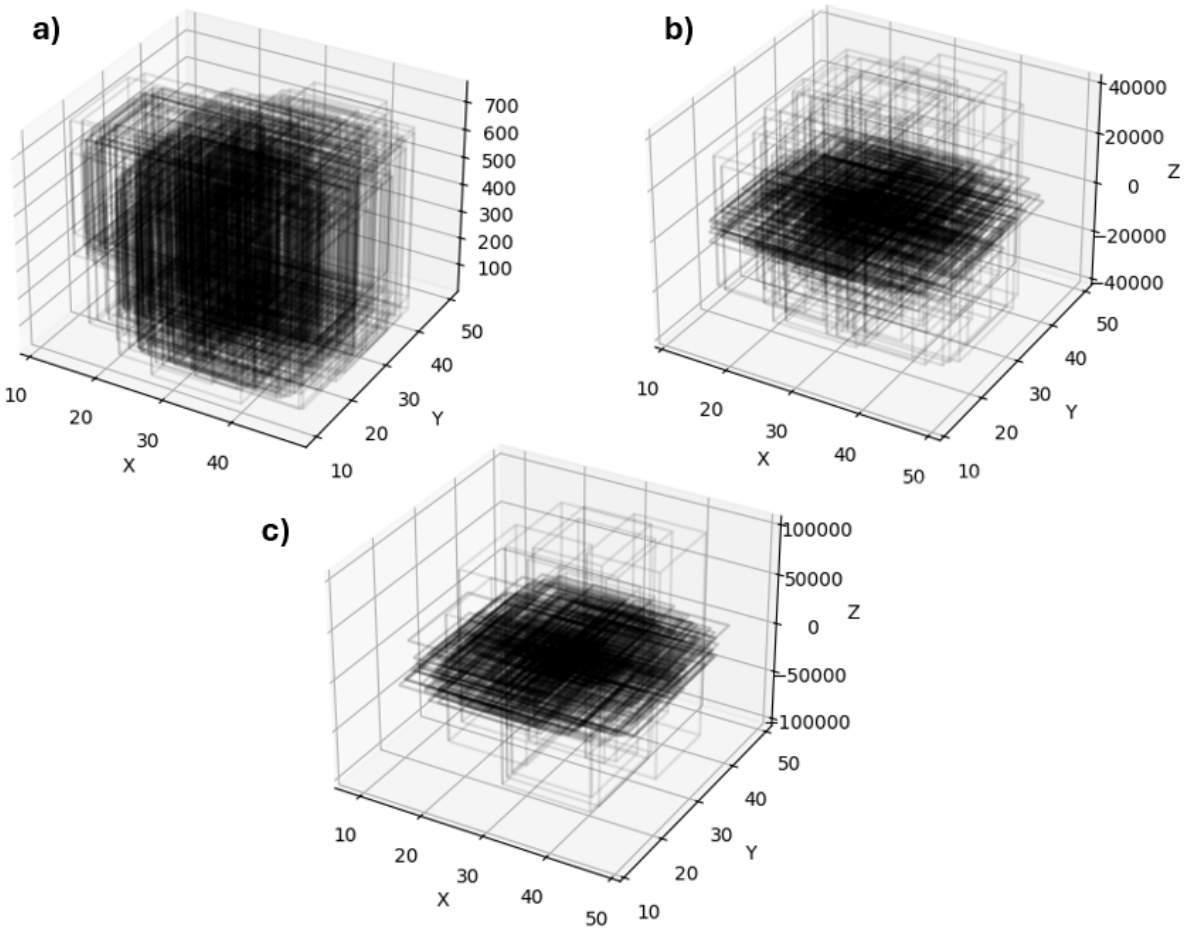
2: **Se Cenário 5** Do conjunto de dados $\{\overline{y}_i, \overline{x}_{i1}, \overline{x}_{i2}\}$ $(i = 1, \ldots, n)$ em ordem decrescente, obtenha o primeiro elemento do conjunto de dados do ponto médio classificado. Calcule valores discrepantes do limite superior por $\overline{y}_i = \overline{y}_i + 9S_{\overline{y}}$ com $(i = 1, ..., k)$ onde $S_{\underline{y}}$ é o desvio padrão do conjunto $(\overline{y}_i, ..., \overline{y}_n)$.

3: **Se Cenário 6** Dos conjuntos de dados $\{\overline{y}_i, \overline{x}_{i1}, \overline{x}_{i2}\}$ $(i = 1, \ldots, n)$ e $\{\overline{y}_i, \overline{x}_{i1}, \overline{x}_{i2}\}$ em ordem decrescente, obtenha valores discrepantes inferiores e superiores simultaneamente como nos cenários 4 e 5.

4.3.0.1 Conjuntos de dados sintéticos com outliers para cenários de 1 a 3

A figura 3 ilustra os conjuntos de dados de intervalo baseados nos cenários 1 a 3 com 10% outliers e tamanho de dados 300 em R^3.

Figura 3 – Cenário 1 (a), Cenário 2 (b) e Cenário 3 (c).

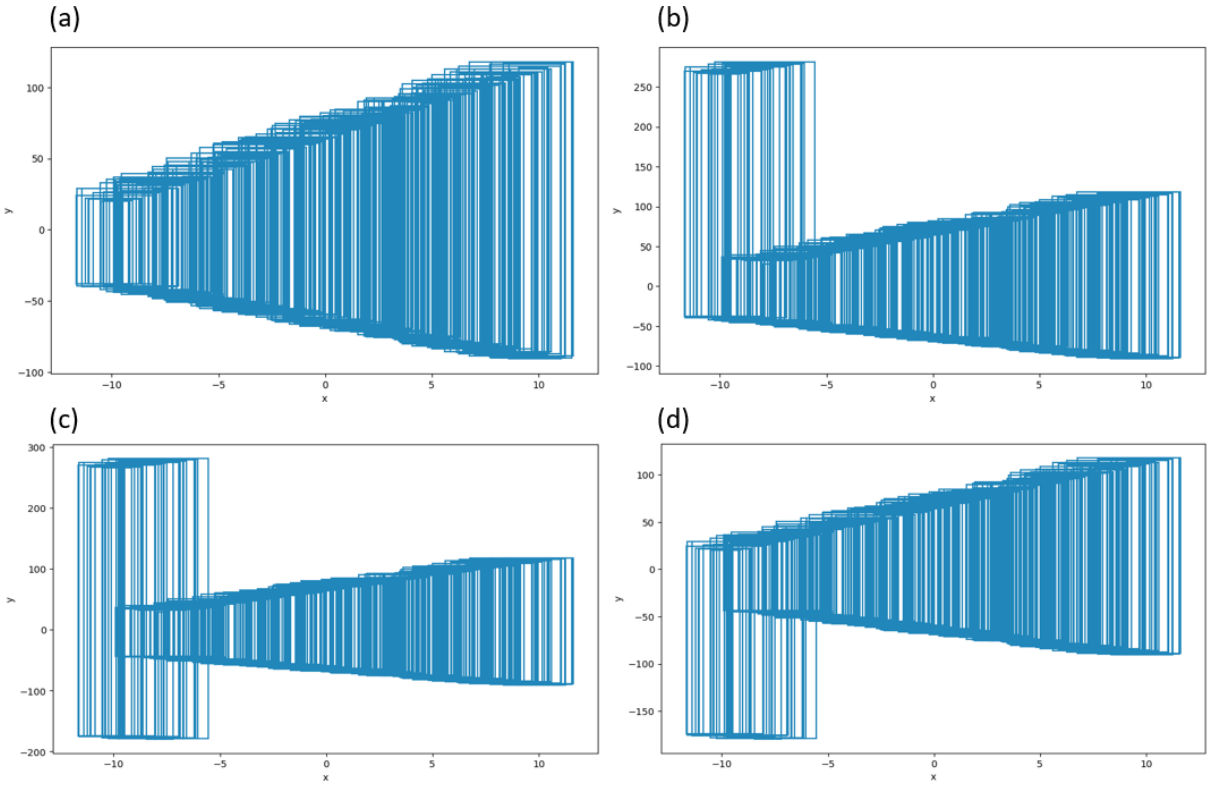


Fonte: O autor (2024)

4.3.0.2 Conjuntos de dados sintéticos com outliers para cenários de 4 a 6

A figura (4) exibe conjuntos de dados de intervalo baseados nos cenários 4 a 6 com 10% de outliers e tamanho de dados 400 em R^2.

Figura 4 – (a) Conjunto de dados intervalares com dependência linear entre variáveis de resposta e preditoras. Conjunto de dados de intervalo com 10% de outliers para limites inferior (d) e superior (b). (c) Conjunto de dados de intervalo com 10% para limites inferior e superior.



Fonte: O autor (2024)

4.3.1 RESULTADOS PARA DADOS SINTÉTICOS

Na avaliação experimental com dados sintéticos, adotamos: tamanho da amostra $n = 400$ e 75% e 25% para os conjuntos de dados de treinamento e teste, respectivamente. Diferentes percentuais de outliers são considerados para cada cenário de dados: 2%, 5% e 10%. O método proposto, chamado SO-IRR, neste artigo é comparado com diferentes métodos da literatura de SDA. Eles são: IRR (FAGUNDES; SOUZA; CYSNEIROS, 2013), iETKRR em relação à representação dos limites inferior e superior, bem como o ponto médio e metade do intervalo, aqui chamado de ieTKRRcr (NETO; CARVALHO, 2018) e LN-IRR (ZHAO; WANG; WANG, 2023). Os resultados com base no intervalo de confiança bootstrap com nível de significância de 5% são apresentados nas Tabelas 2, 3 e 4 de acordo com MMRE, MMRE:L e MMRE:U, respectivamente.

Em geral, para os cenários 1 a 3, podemos observar que SO-IRR, IRR e LN-IRR têm desempenho similar em termos das métricas usadas e percentagem de outliers. Mas para os

cenários 4, 5 e 6, nos quais há dependência de ambos os limites de resposta para pontos internos da variável regressora, o método proposto supera os outros em termos de todas as métricas e percentagens de outliers.

Tabela 2 – Resultados dos dados sintéticos para a métrica MMRE.

Cenário 1					
Outliers	**IRR**	**iETKRR**	**iETKRRcr**	**LN-IRR**	**SO-IRR**
2%	**[0.020 , 0.021]**	[0.058 , 0.062]	[0.055 , 0.061]	**[0.020 , 0.021]**	**[0.020 , 0.021]**
5%	**[0.021 , 0.023]**	[0.131 , 0.140]	[0.131 , 0.141]	**[0.022 , 0.024]**	**[0.021 , 0.023]**
10%	**[0.021 , 0.023]**	[0.231 , 0.240]	[0.234 , 0.250]	[0.028 , 0.030]	**[0.022 , 0.023]**
Cenário 2					
2%	**[0.020 , 0.021]**	[0.179 , 0.204]	[0.055 , 0.073]	**[0.020 , 0.021]**	**[0.020 , 0.022]**
5%	**[0.022 , 0.025]**	[0.476 , 0.535]	[0.104 , 0.148]	**[0.022 , 0.025]**	**[0.022 , 0.025]**
10%	**[0.022 , 0.024]**	[0.895 , 1.057]	[0.339 , 0.438]	**[0.022 , 0.024]**	**[0.022 , 0.024]**
Cenário 3					
2%	**[0.021 , 0.024]**	[0.080 , 0.091]	[0.054 , 0.059]	**[0.022 , 0.024]**	**[0.021 , 0.024]**
5%	**[0.021 , 0.022]**	[0.198 , 0.214]	[0.129 , 0.138]	**[0.021 , 0.023]**	**[0.020 , 0.022]**
10%	**[0.022 , 0.033]**	[0.363 , 0.388]	[0.252 , 0.274]	**[0.030 , 0.042]**	**[0.022 , 0.034]**
Cenário 4					
2%	[0.138 , 0.158]	[0.046 , 0.051]	[0.146 , 0.164]	[0.134 , 0.154]	**[0.001 , 0.008]**
5%	[0.132 , 0.152]	[0.111 , 0.125]	[0.176 , 0.197]	[0.129 , 0.148]	**[0.000 , 0.000]**
10%	[0.149 , 0.172]	[0.216 , 0.249]	[0.284 , 0.315]	[0.147 , 0.169]	**[0.001 , 0.010]**
Cenário 5					
2%	[0.143 , 0.164]	[0.044 , 0.050]	[0.150 , 0.170]	[0.139 , 0.159]	**[0.000 , 0.003]**
5%	[0.134 , 0.153]	[0.106 , 0.121]	[0.176 , 0.194]	[0.131 , 0.150]	**[0.000 , 0.003]**
10%	[0.154 , 0.177]	[0.240 , 0.270]	[0.282 , 0.309]	[0.151 , 0.172]	**[0.000 , 0.005]**
Cenário 6					
2%	[0.141 , 0.163]	[0.064 , 0.072]	[0.151 , 0.169]	[0.138 , 0.158]	**[0.000 , 0.000]**
5%	[0.146 , 0.166]	[0.191 , 0.211]	[0.194 , 0.219]	[0.144 , 0.165]	**[0.000 , 0.000]**
10%	[0.147 , 0.167]	[0.418 , 0.459]	[0.312 , 0.344]	[0.149 , 0.170]	**[0.000 , 0.003]**

Fonte: O autor (2024)

Tabela 3 – Resultados dos dados sintéticos para a métrica MMRE:L.

Outliers	IRR	iETKRR	iETKRRcr	LN-IRR	SO-IRR
Cenário 1					
2%	**[0.020 , 0.022]**	[0.054 , 0.058]	[0.055 , 0.061]	**[0.020 , 0.022]**	**[0.020 , 0.022]**
5%	**[0.021 , 0.023]**	[0.122 , 0.132]	[0.123 , 0.134]	**[0.022 , 0.023]**	**[0.021 , 0.023]**
10%	**[0.021 , 0.022]**	[0.235 , 0.248]	[0.247 , 0.263]	[0.028 , 0.030]	**[0.021 , 0.022]**
Cenário 2					
2%	**[0.023 , 0.026]**	[0.198 , 0.221]	[0.050 , 0.060]	**[0.023 , 0.026]**	**[0.023 , 0.026]**
5%	**[0.021 , 0.022]**	[0.500 , 0.566]	[0.133 , 0.175]	**[0.021 , 0.022]**	**[0.021 , 0.022]**
10%	**[0.021 , 0.023]**	[1.048 , 1.144]	[0.291 , 0.385]	**[0.021 , 0.023]**	**[0.022 , 0.023]**
Cenário 3					
2%	**[0.025 , 0.032]**	[0.062 , 0.076]	[0.062 , 0.069]	**[0.026 , 0.033]**	**[0.025 , 0.033]**
5%	**[0.023 , 0.026]**	[0.163 , 0.188]	[0.126 , 0.135]	**[0.024 , 0.027]**	**[0.023 , 0.026]**
10%	**[0.025 , 0.026]**	[0.288 , 0.335]	[0.255 , 0.272]	[0.033 , 0.035]	**[0.025 , 0.027]**
Cenário 4					
2%	[0.143 , 0.164]	[0.062 , 0.074]	[0.151 , 0.174]	[0.140 , 0.159]	**[0.000 , 0.000]**
5%	[0.144 , 0.166]	[0.178 , 0.202]	[0.193 , 0.222]	[0.140 , 0.161]	**[0.000 , 0.006]**
10%	[0.145 , 0.169]	[0.407 , 0.461]	[0.336 , 0.392]	[0.142 , 0.166]	**[0.000 , 0.000]**
Cenário 5					
2%	[0.142 , 0.164]	[0.024 , 0.028]	[0.148 , 0.166]	[0.139 , 0.159]	**[0.000 , 0.000]**
5%	[0.150 , 0.172]	[0.023 , 0.028]	[0.168 , 0.191]	[0.146 , 0.167]	**[0.000 , 0.012]**
10%	[0.152 , 0.177]	[0.023 , 0.028]	[0.210 , 0.239]	[0.146 , 0.169]	**[0.000 , 0.006]**
Cenário 6					
2%	[0.150 , 0.172]	[0.066 , 0.076]	[0.159 , 0.184]	[0.147 , 0.167]	**[0.000 , 0.000]**
5%	[0.142 , 0.162]	[0.198 , 0.228]	[0.189 , 0.216]	[0.140 , 0.160]	**[0.000 , 0.000]**
10%	[0.143 , 0.165]	[0.419 , 0.474]	[0.308 , 0.354]	[0.145 , 0.168]	**[0.000 , 0.000]**

Fonte: O autor (2024)

Tabela 4 – Resultados dos dados sintéticos para a métrica MMRE:U.

Cenário 1					
Outliers	**IRR**	**iETKRR**	**iETKRRcr**	**LN-IRR**	**SO-IRR**
2%	**[0.020 , 0.021]**	[0.058 , 0.063]	[0.055 , 0.060]	**[0.020 , 0.021]**	**[0.020 , 0.022]**
5%	**[0.021 , 0.023]**	[0.130 , 0.139]	[0.134 , 0.144]	**[0.022 , 0.024]**	**[0.021 , 0.023]**
10%	**[0.020 , 0.022]**	[0.239 , 0.250]	[0.230 , 0.246]	[0.027 , 0.030]	**[0.020 , 0.023]**
Cenário 2					
2%	**[0.019 , 0.020]**	[0.202 , 0.227]	[0.049 , 0.064]	**[0.019 , 0.020]**	**[0.019 , 0.020]**
5%	**[0.019 , 0.020]**	[0.480 , 0.546]	[0.133 , 0.179]	**[0.019 , 0.020]**	**[0.019 , 0.020]**
10%	**[0.020 , 0.022]**	[1.032 , 1.123]	[0.302 , 0.393]	**[0.020 , 0.022]**	**[0.020 , 0.022]**
Cenário 3					
2%	**[0.019 , 0.021]**	[0.093 , 0.103]	[0.060 , 0.066]	**[0.020 , 0.021]**	**[0.020 , 0.021]**
5%	**[0.021 , 0.023]**	[0.199 , 0.233]	[0.141 , 0.161]	**[0.022 , 0.023]**	**[0.021 , 0.022]**
10%	**[0.019 , 0.021]**	[0.438 , 0.476]	[0.265 , 0.286]	[0.026 , 0.028]	**[0.020 , 0.021]**
Cenário 4					
2%	[0.136 , 0.158]	[0.022 , 0.027]	[0.143 , 0.160]	[0.133 , 0.152]	**[0.000 , 0.000]**
5%	[0.140 , 0.160]	[0.026 , 0.030]	[0.169 , 0.191]	[0.137 , 0.156]	**[0.000 , 0.006]**
10%	[0.144 , 0.169]	[0.025 , 0.030]	[0.195 , 0.230]	[0.138 , 0.162]	**[0.000 , 0.000]**
Cenário 5					
2%	[0.134 , 0.153]	[0.064 , 0.075]	[0.143 , 0.163]	[0.132 , 0.150]	**[0.000 , 0.000]**
5%	[0.147 , 0.169]	[0.185 , 0.213]	[0.209 , 0.239]	[0.143 , 0.165]	**[0.000 , 0.000]**
10%	[0.145 , 0.166]	[0.396 , 0.463]	[0.310 , 0.358]	[0.145 , 0.165]	**[0.000 , 0.000]**
Cenário 6					
2%	[0.136 , 0.157]	[0.064 , 0.076]	[0.144 , 0.164]	[0.132 , 0.152]	**[0.000 , 0.006]**
5%	[0.148 , 0.166]	[0.178 , 0.207]	[0.189 , 0.220]	[0.146 , 0.165]	**[0.000 , 0.006]**
10%	[0.142 , 0.164]	[0.411 , 0.472]	[0.304 , 0.347]	[0.146 , 0.169]	**[0.000 , 0.000]**

Fonte: O autor (2024)

4.4 DADOS REAIS

Ao avaliar modelos com dados reais, diferentes técnicas de validação serão usadas dependendo do tamanho do conjunto de dados. Para conjuntos de dados com menos de 100 observações, a técnica leave-one-out é aplicada. Para conjuntos de dados com mais de 100 observações, a técnica de validação cruzada 5-fold é usada. Este esquema de validação cruzada é repetido 20 vezes. As métricas $MMRE$, $MMRE : L$ e $MMRE : U$ são representadas em termos de sua média e desvio padrão. O teste de hipótese estatística de Friedman é aplicado e o pós-teste de Nemenyi é empregado. Seis conjuntos de dados reais intervalares são considerados: Cardiologia, Carros, Cogumelos, Futebol, Raças de Cães e Clima.

O conjunto de dados intervalares de Cardiologia consiste em 59 pacientes descritos por três variáveis intervalares. Entre elas, estão duas variáveis de regressão independentes, Pressão Arterial Sistólica e Pressão Arterial Diastólica, e uma variável de resposta, Frequência Cardíaca. O conjunto de dados de Carros consiste em 33 modelos descritos por três variáveis preditoras intervalares, que são Velocidade Máxima, Deslocamento do Motor e Preço como variável de

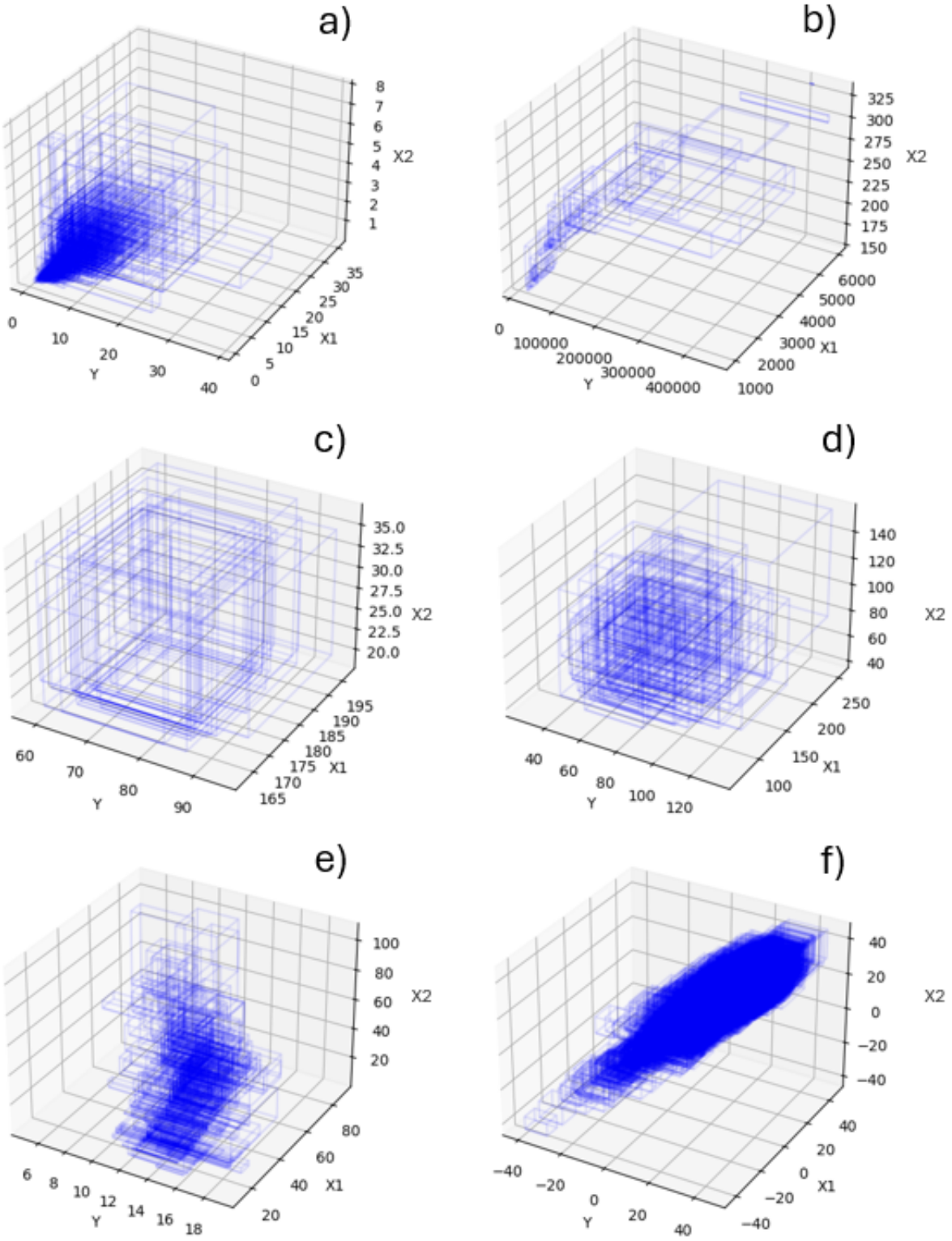
resposta. Este conjunto de dados foi usado em (NETO; CARVALHO, 2008) e (SILVA; BRITO, 2006).

O conjunto de dados de Cogumelos contém medidas de três características de 100 espécies de cogumelos, que são membros do gênero Agaricus. Todas as medidas no conjunto de dados são intervalares, extraídas do Fungi of California Species Index. Comprimento do Estipe e Espessura do Estipe são variáveis preditoras, respectivamente, e Largura do Píleo é a variável de resposta. Ao todo, 274 observações são registradas no conjunto de dados (XU, 2010).

O conjunto de dados de Futebol fornece informações sobre jogadores de futebol profissional em 20 times franceses. Neste conjunto de dados, Altura e Idade foram consideradas como variáveis preditoras e Peso é a variável dependente intervalar (FAGUNDES; SOUZA; CYSNEIROS, 2013). O conjunto de dados de Raças de Cães contém informações sobre 270 raças e foi extraído do site do American Kennel Club. Este conjunto de dados contém 20 características de diferentes tipos, incluindo dados textuais, numéricos e categóricos. Altura, Peso e Expectativa de Vida são usados para o problema de regressão em (CARVALHO; NETO; ROSENDO, 2021).

O conjunto de dados de Clima usado em (CARVALHO; NETO; ROSENDO, 2021) apresenta Observações Meteorológicas Oficiais, Previsões Meteorológicas e Informações Climatológicas dos Serviços Meteorológicos e Hidrológicos Nacionais em todo o mundo. Tem 5660 observações com o objetivo de prever a temperatura ([min, max]) no último mês da estação com base nos dois meses anteriores. Assim, a variável de resposta é o intervalo de temperatura da última estação e duas variáveis explicativas representam os intervalos de temperatura dos dois meses anteriores. A Figura 5 exibe esses conjuntos de dados reais intervalares.

Figura 5 – Dados intervalares a) Cogumelos b) Carros c) Futebol d) Cardiologia e) Raças de cães e f) Clima.



Fonte: O autor (2024)

4.5 ANALISE DE DESEMPENHO PARA DADOS REAIS

A Tabela 5 mostra os resultados para MMRE, MMRE:L e MMRE:U e os conjuntos de dados reais. A partir desta tabela, podemos observar que, em geral, o método proposto é uma boa opção em comparação com os métodos da literatura de SDA em termos das métricas adotadas. A Tabela 6 apresenta os valores-p para o teste de Friedman. Com relação a esses valores-p, podemos dizer que não há evidências suficientes para rejeitar a hipótese nula para os conjuntos de dados de futebol e cardiologia. As Figuras 6, 7, 8 e 8 mostram o mapa de calor do pós-teste de Nemenyi para os conjuntos de dados em que o valor-p de Friedman é menor ou igual a 0,05.

Tabela 5 – Resultados para conjuntos de dados reais

	Cogumelos			**Carros**		
	MMRE	**MMRE:L**	**MMRE:U**	**MMRE**	**MMRE:L**	**MMRE:U**
IRR	0.404 (0.128)	0.429 (0.133)	0.380 (0.132)	0.420 (0.398)	0.504 (0.474)	0.336 (0.385)
iETKRR	0.399 (0.130)	0.430 (0.140)	**0.369 (0.130)**	0.436 (0.490)	0.502 (0.759)	0.370 (0.301)
iETKRR_cr	0.502 (0.144)	0.580 (0.167)	0.423 (0.147)	0.443 (0.431)	0.494 (0.622)	0.391 (0.346)
LN-IRR	0.456 (0.119)	0.497 (0.120)	0.415 (0.136)	0.402 (0.402)	0.503 (0.510)	0.301 (0.338)
SO-IRR	**0.388 (0.124)**	**0.402 (0.124)**	0.373 (0.133)	**0.257 (0.195)**	**0.232 (0.206)**	**0.282 (0.248)**
	Futebol			**Cardiologia**		
IRR	0.027 (0.019)	0.030 (0.023)	0.024 (0.028)	**0.154 (0.112)**	0.166 (0.172)	0.143 (0.111)
iETKRR	0.031 (0.016)	0.039 (0.032)	0.022 (0.025)	**0.154 (0.124)**	**0.162 (0.190)**	0.146 (0.120)
iETKRR_cr	0.030 (0.020)	0.032 (0.027)	0.028 (0.028)	0.155 (0.119)	0.170 (0.179)	**0.140 (0.120)**
LN-IRR	0.027 (0.019)	0.029 (0.022)	0.024 (0.027)	0.155 (0.117)	0.169 (0.183)	0.142 (0.106)
SO-IRR	**0.025 (0.019)**	**0.028 (0.025)**	**0.021 (0.026)**	0.158 (0.127)	0.169 (0.196)	0.147 (0.125)
	Raças de cães			**Clima**		
IRR	**0.090 (0.011)**	0.097 (0.014)	0.083 (0.010)	0.254 (0.028)	**0.254 (0.031)**	0.253 (0.041)
iETKRR	0.094 (0.013)	0.102 (0.020)	0.087 (0.010)	0.280 (0.027)	0.325 (0.036)	**0.235 (0.037)**
iETKRR_cr	0.091 (0.012)	0.097 (0.015)	0.085 (0.011)	0.297 (0.032)	0.279 (0.032)	0.316 (0.049)
LN-IRR	**0.090 (0.011)**	0.097 (0.014)	**0.082** (0.010)	0.322 (0.030)	0.341 (0.034)	0.302 (0.047)
SO-IRR	**0.090 (0.011)**	**0.096** (0.014)	0.083 (0.010)	**0.244 (0.027)**	**0.254 (0.031)**	**0.235 (0.039)**

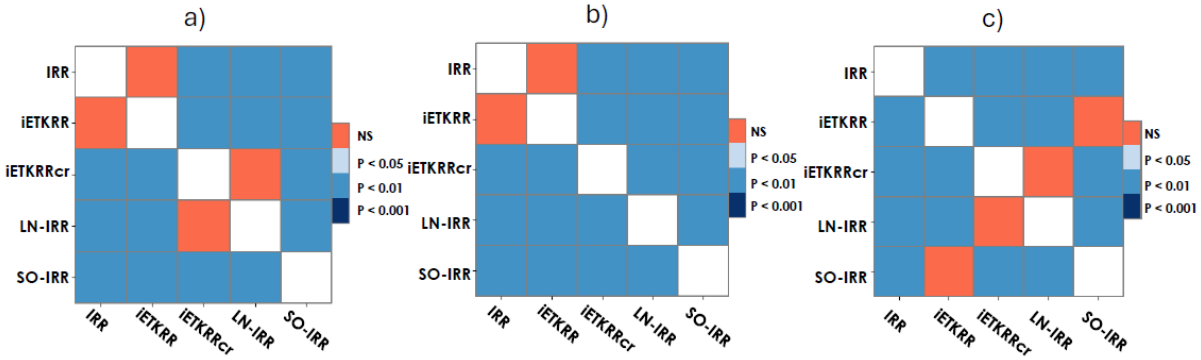
Fonte: O autor (2024)

Tabela 6 – P Valor para o teste estatístico de Friedman

	MMRE	**MMRE:L**	**MMRE:U**
Cogumelos	0	0	0
Carros	0	0.011	0.211
Futebol	0.112	0.256	0.081
Cardiologia	0.929	0.231	0.974
Raças de cães	0	0	0
Clima	0	0	0

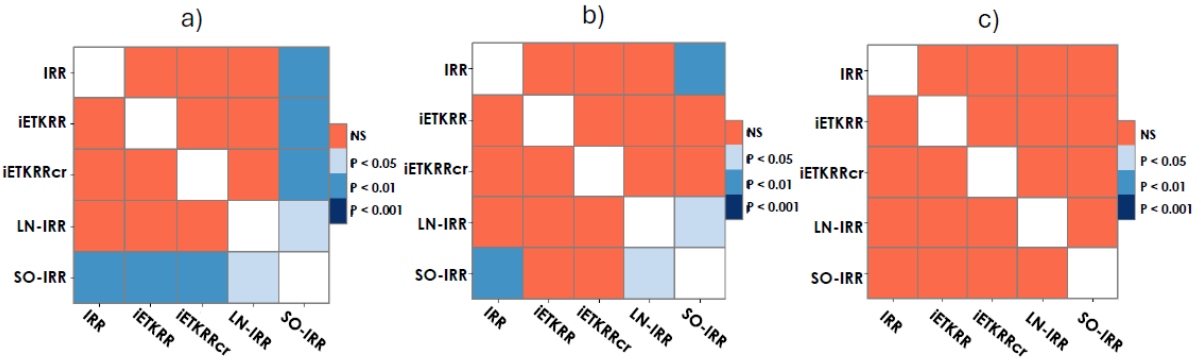
Fonte: O autor (2024)

Figura 6 – Pós-teste de Nemenyi para o conjunto de dados de cogumelos. MMRE (a), MMRE:L (b) e MMRE:U (c).



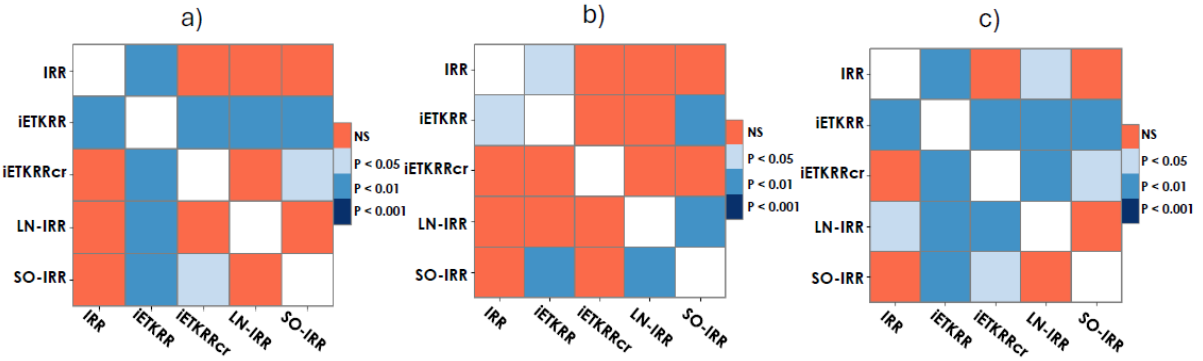
Fonte: O autor (2024)

Figura 7 – Pós-teste de Nemenyi para o conjunto de dados de carros. MMRE (a), MMRE:L (b) e MMRE:U (c).



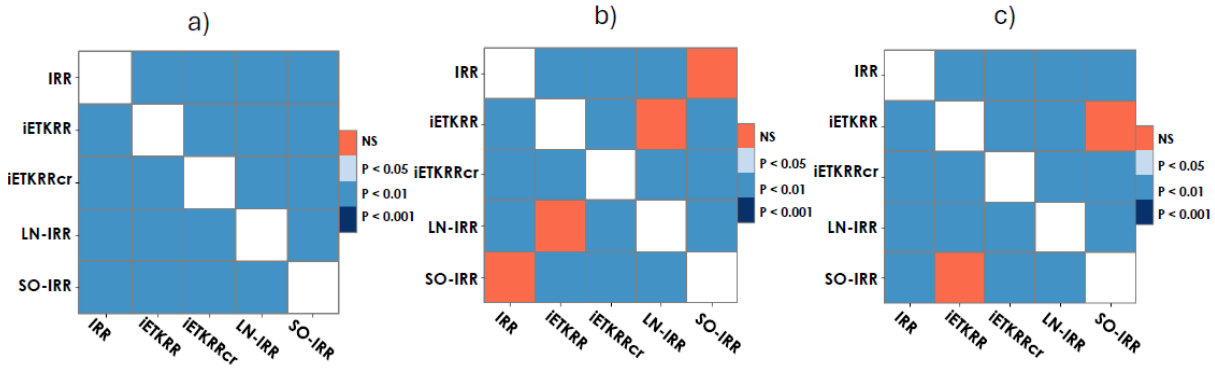
Fonte: O autor (2024)

Figura 8 – Pós-teste de Nemenyi para o conjunto de dados de Raças de cães. MMRE (a), MMRE:L (b) e MMRE:U (c).



Fonte: O autor (2024)

Figura 9 – Pós-teste de Nemenyi para o conjunto de dados de clima. MMRE (a), MMRE:L (b) e MMRE:U (c).



Fonte: O autor (2024)

Os resultados para o conjunto de dados de Cogumelos na Figura 6 indicam que o modelo proposto apresentou a menor média para as métricas $MMRE$ e $MMRE:L$. Para a métrica $MMRE:U$, o método proposto é tão bom quanto o iETKRR, sendo ambos as melhores opções entre todos os métodos. Para os dados de Carros, o modelo proposto apresentou a menor média para todas as métricas e, de acordo com o pós-teste de Nemenyi par a par, 7, o SO-IRR superou os outros métodos para as métricas $MMRE$ e $MMRE:L$. No entanto, para a métrica $MMRE:U$, não há evidências suficientes para rejeitar a hipótese nula.

Para os dados de Raças de Cães, o SO-IRR é tão bom quanto o LN-IRR e o IRR em termos de $MMRE$ e $MMRE:U$. Em relação ao $MMRE:L$, o SO-IRR é tão bom quanto o LN-IRR. Nesse contexto, SO-IRR, IRR e LN-IRR são as melhores opções e a Figura 8 exibe esses resultados para os dados de Raças de Cães. Para o conjunto de dados de Clima, podemos dizer que o método proposto é superior aos outros em termos de $MMRE$, tão bom quanto o IRR em termos de $MMRE:L$ e tão bom quanto o iETKRR em termos de $MMRE:U$. A Figura 9 apresenta esses resultados para o conjunto de dados de clima.

4.6 ANALISE DE OUTLIERS

Esta seção apresenta uma avaliação sobre os candidatos a outliers intervalares em conjuntos de dados reais de acordo com as Equações (3.16) e (3.17). Assim, um intervalo de resposta é um candidato a outlier se $|\underline{\Delta_i}| \geq \tau$ e/ou $|\overline{\Delta_i}| \geq \tau$. Aqui, definimos $\tau = 2$, pois é constantemente usado na literatura para indicar resíduo grande. A partir desta avaliação, podemos observar que não há candidatos a outliers de limite inferior ou superior para os conjuntos de

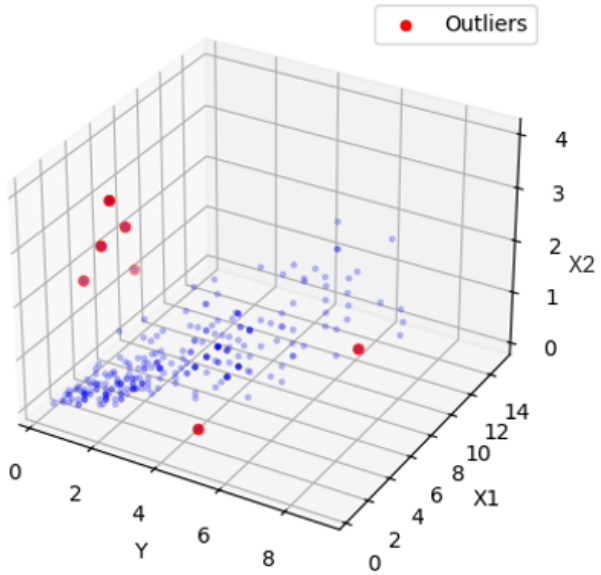
dados de Carros, Cardiologia e Futebol.

Por outro lado, há 7 intervalos candidatos a outliers baseados nos limites inferiores para o conjunto de dados de Cogumelos. De acordo com a definição de outlier intervalar na Equação (3.16), temos: $|\underline{\Delta}_{18}| = 2.1,\ |\underline{\Delta}_{104}| = 3.9,\ |\underline{\Delta}_{138}| = 3.8,\ |\underline{\Delta}_{187}| = 3.5,\ |\underline{\Delta}_{206}| = 2.0,\ |\underline{\Delta}_{239}| = 4.4,\ |\underline{\Delta}_{262}| = 3.1$. A Figura 10 ilustra esses candidatos.

Para o conjunto de dados de clima, há 9 intervalos candidatos a outliers para o limite inferior e, de acordo com a definição na Equação (3.16), temos: $|\underline{\Delta}_{541}| = 2.2,\ |\underline{\Delta}_{1157}| = 2.0,\ |\underline{\Delta}_{1714}| = 2.3,\ |\underline{\Delta}_{2460}| = 2.0,\ |\underline{\Delta}_{2520}| = 2.0, |\underline{\Delta}_{25844}| = 2.1, |\underline{\Delta}_{2683}| = 2.2, |\underline{\Delta}_{2876}| = 2.0, |\underline{\Delta}_{3248}| = 2.0$. Além disso, há 2 intervalos candidatos a outliers para o limite superior e, de acordo com a definição na Equação (3.17), temos: $|\overline{\Delta}_{2398}| = 2.2$ e $|\overline{\Delta}_{4579}| = 2.1$. Seguindo a definição de outliers intervalares baseados nos limites inferior e superior, temos 4 intervalos candidatos que são: ($|\underline{\Delta}_{1161}| = 2.5$ e $|\overline{\Delta}_{1161}| = 2.0$); ($|\underline{\Delta}_{1170}| = 3.0$ e $|\overline{\Delta}_{1170}| = 2.3$); ($|\underline{\Delta}_{2571}| = 2.8$ e $|\overline{\Delta}_{2571}| = 2.0$); ($|\underline{\Delta}_{2575}| = 2.7$, e $|\overline{\Delta}_{2575}| = 2.4$). A Figura 11 ilustra esses candidatos.

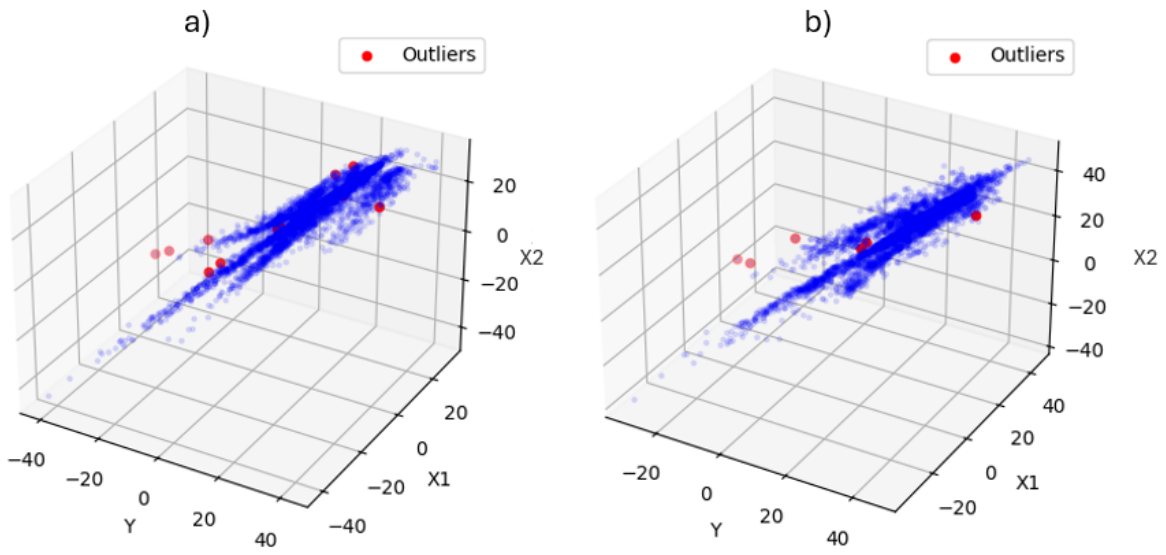
Para os dados de Raças de Cães, há 2 intervalos candidatos a outliers para o limite inferior de acordo com $|\underline{\Delta}_{86}| = 2.6$ e $|\underline{\Delta}_{104}| = 2.4$ e 1 para o limite superior de acordo com $|\overline{\Delta}_{183}| = 2.0$. A Figura 12 ilustra esses candidatos.

Figura 10 – Candidatos a outliers para conjunto de dados de Cogumelo



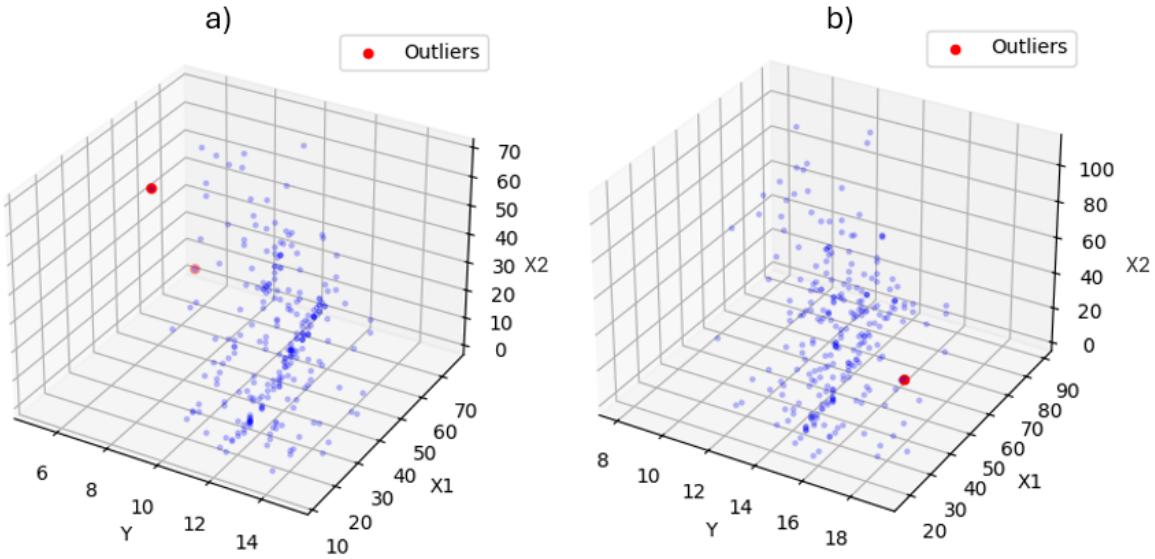
Fonte: O autor (2024)

Figura 11 – Candidatos a outliers para conjunto de dados de Clima a) limite inferior e b) limite superior.



Fonte: O autor (2024)

Figura 12 – Candidatos a outliers para conjunto de dados de Raças de Cão a) limite inferior e b) limite superior.



Fonte: O autor (2024)

5 CONCLUSÃO

Este trabalho propôs um método de regressão linear robusta para dados intervalares, no qual os intervalos são representados por diferentes pontos de referência que são automaticamente extraídos como os melhores pontos de referência das variáveis regressoras. Esta forma de representação permite que o próprio modelo encontre os melhores ajustes que expliquem o comportamento dos limites da variável de resposta intervalar. Por essa razão, chamamos o modelo robusto de auto-organizado. Além disso, este modelo é menos sensível na presença de outliers intervalares. Portanto, este artigo também propõe a definição de outliers para os limites inferiores e superiores.

Uma avaliação experimental com dados sintéticos é realizada utilizando três métricas baseadas na magnitude média do erro relativo. Diferentes configurações de dados sintéticos são consideradas e intervalos de confiança Bootstrap para as métricas são obtidos. Seis cenários de configurações de dados são adotados. Três cenários de outliers são definidos em termos do centro e do alcance dos intervalos, de acordo com a literatura SDA para modelos robustos, e três cenários são definidos de forma que os limites da variável de resposta dependam de pontos internos da variável preditora. Os resultados destacam que o modelo proposto é tão bom quanto os modelos da literatura SDA, mas supera esses modelos quando há dependência de ambos os limites da resposta em relação a pontos internos da variável regressora, em termos de todas as métricas e percentuais de outliers.

Quanto aos dados intervalares reais, seis conjuntos comumente utilizados na literatura SDA são adotados e as métricas são estimadas com base na técnica de validação cruzada 5-fold no contexto de Monte Carlo com 20 repetições. A análise de desempenho baseada em teste de hipótese indicou que não há evidência para dizer que os modelos avaliados são diferentes em termos de magnitude média do erro relativo com 95% de significância para os conjuntos de dados de Futebol e Cardiologia. No entanto, para os conjuntos de dados de Cogumelos, Carros e Clima, o modelo apresentado neste artigo é a melhor opção. Para o conjunto de dados de Raças de Cães, o modelo introduzido é um dos melhores. A análise de outliers mostrou que, de acordo com as definições propostas neste artigo, os conjuntos de dados de Clima, Raças de Cães e Cogumelos possuem candidatos a outliers intervalares.

5.1 TRABALHOS FUTUROS

O desenvolvimento de aplicações que resolva problemas reais trazem uma maior visibilidade ao método proposto. Dessa forma, se torna interessante a construção de uma aplicação usando o método SO-IRR.

Além disso, no campo da regressão robusta para dados de intervalo, é possível propor um método utilizando algumas abordagens conhecidas como Kernel o logica difusa, visando obter uma performance ainda melhor.

BIBLIOGRAFIA

AL-ASADI, M. Interval-valued data analysis: a review. *Artificial Intelligence Studies*, v. 5, n. 2, p. 47–55, 2022.

BEATON, A. E.; TUKEY, J. W. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, Taylor & Francis, v. 16, n. 2, p. 147–185, 1974.

BILLARD, L.; DIDAY, E. Regression analysis for interval-valued data. In: *Data analysis, classification, and related methods*. [S.l.]: Springer, 2000. p. 369–374.

BILLARD, L.; DIDAY, E. Symbolic regression analysis. In: *Classification, clustering, and data analysis: recent advances and applications*. [S.l.]: Springer, 2002. p. 281–288.

BOCK, H.-H.; DIDAY, E. *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data*. [S.l.]: Springer Science & Business Media, 1999.

BRITO, P.; DIAS, S. *Analysis of distributional data*. [S.l.]: CRC Press, 2022.

CARVALHO, F. d. A. T. de; NETO, E. d. A. L.; ROSENDO, U. d. N. Interval joint robust regression method. *Neurocomputing*, Elsevier, v. 465, p. 265–288, 2021.

DIDAY, E.; NOIRHOMME-FRAITURE, M. *Symbolic data analysis and the SODAS software*. [S.l.]: John Wiley & Sons, 2008.

FAGUNDES, R. A.; SOUZA, R. M. D.; CYSNEIROS, F. J. A. Robust regression with application to symbolic interval data. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 26, n. 1, p. 564–573, 2013.

FERRARO, M. B.; GIORDANI, P. A proposal of robust regression for random fuzzy sets. In: SPRINGER. *Synergies of Soft Computing and Statistics for Intelligent Data Analysis*. [S.l.], 2013. p. 115–123.

FREITAS, W. W.; SOUZA, R. M. de; AMARAL, G. J.; MORAES, R. M. de. Regression applied to symbolic interval-spatial data. *Applied Intelligence*, Springer, v. 54, n. 2, p. 1545–1565, 2024.

GIORDANI, P. Lasso-constrained regression analysis for interval-valued data. *Advances in Data Analysis and Classification*, Springer, v. 9, n. 1, p. 5–19, 2015.

HAO, P.; GUO, J. Constrained center and range joint model for interval-valued symbolic data regression. *Computational Statistics & Data Analysis*, Elsevier, v. 116, p. 106–138, 2017.

LAWSON, C. L.; HANSON, R. J. Linear least squares with linear inequality constraints. *Solving least squares problems*, Prentice-hall Englewood Cliffs, NJ, p. 158–173, 1974.

LEITHOLD, L. The calculus with analytic geometry. *(No Title)*, 1972.

MCCREA, W. H. *Analytical geometry of three dimensions*. [S.l.]: Courier Corporation, 2012.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Introduction to linear regression analysis*. [S.l.]: John Wiley & Sons, 2021.

MOORE, R. E. *Interval analysis*. [S.l.]: Prentice-Hall, 1966.

NAJAFABADI, M. M.; VILLANUSTRE, F.; KHOSHGOFTAAR, T. M.; SELIYA, N.; WALD, R.; MUHAREMAGIC, E. Deep learning applications and challenges in big data analytics. *Journal of big data*, SpringerOpen, v. 2, n. 1, p. 1–21, 2015.

NETO, E. A. L.; CARVALHO, F. A. T. D. Centre and range method for fitting a linear regression model to symbolic interval data. *Computational Statistics & Data Analysis*, Elsevier, v. 52, n. 3, p. 1500–1515, 2008.

NETO, E. A. L.; CARVALHO, F. A. T. D. Constrained linear regression models for symbolic interval-valued variables. *Computational Statistics & Data Analysis*, Elsevier, v. 54, n. 2, p. 333–347, 2010.

NETO, E. A. L.; CARVALHO, F. A. T. de. An exponential-type kernel robust regression model for interval-valued variables. *Information Sciences*, Elsevier, v. 454, p. 419–442, 2018.

REYNOLDS, F. J. *Using Symbolic Data Analysis to Detect Fraud, Waste, and Abuse in Healthcare Insurance Claims Data*. Tese (Doutorado) — Auburn University, 2020.

SILVA, A. P. D.; BRITO, P. Linear discriminant analysis for interval data. *Computational Statistics*, Springer, v. 21, p. 289–308, 2006.

SINOVA, B.; COLUBI, A.; GONZÁLEZ-RODRI, G. et al. Interval arithmetic-based simple linear regression between interval data: Discussion and sensitivity analysis on the choice of the metric. *Information Sciences*, Elsevier, v. 199, p. 109–124, 2012.

SOARES, Y. M.; FAGUNDES, R. A. Interval quantile regression models based on swarm intelligence. *Applied Soft Computing*, Elsevier, v. 72, p. 474–485, 2018.

SOUZA, L. C.; SOUZA, R. M.; AMARAL, G. J.; FILHO, T. M. S. A parametrized approach for linear regression of interval data. *Knowledge-Based Systems*, Elsevier, v. 131, p. 149–159, 2017.

WANG, H.; GUAN, R.; WU, J. Linear regression of interval-valued data based on complete information in hypercubes. *Journal of Systems Science and Systems Engineering*, Springer, v. 21, n. 4, p. 422–442, 2012.

XU, W. *Symbolic data analysis: interval-valued data regression*. Tese (Doutorado) — University of Georgia Athens, GA, 2010.

ZHAO, Q.; WANG, H.; WANG, S. Robust regression for interval-valued data based on midpoints and log-ranges. *Advances in Data Analysis and Classification*, Springer, v. 17, p. 583–621, 2023.