MODELAGEM DE REGRESSÃO LINEAR COM ARITMÉTICA INTERVALAR MULTIDIMENSIONAL RDM

LUCAS M. TORTELLI[1]; DIRCEU MARASCHIN JR.[2]; ALINE B. LORETO[3]

[1]PPGC-UFPel– lmtortelli@inf.ufpel.edu.br
[2]PPGC-UFPel– dmaraschin@inf.ufpel.edu.br
[3]UFSM – aline.loreto@ufsm.br

1. INTRODUÇÃO

Na computação convencional é notável que os erros são inerentes às operações sobre a aritmética de ponto flutuante, uma vez que esta representa seus valores através de um subconjunto dos números reais. Esta aritmética contém representatividade finita sobre todos os valores que pode assumir, ocasionando deficiências quanto à necessidade de precisão em cálculos numéricos. Não somente decorrente da representaçãom que surgem os erros, também a partir dos resultados intermediários de cada operação, que vão acumulando erros ao decorrer da execução (GOLDBERG,1991).

A proposta sugerida para correção numérica nas máquinas foi concebida por MOORE (2003), conhecida como SIA (do inglês, *Standard Interval Arithmetic*). A SIA representa valores de ponto flutuante através de intervalos [x]. Isto proporciona ferramentas robustas de mensuração da qualidade do resultado, agregando garantias aos valores numéricos obtidos, tanto pela sua representatividade como também por fornecer a inexatidão presente no intervalo (RATSCHEK, 1988).

Análises de regressão visa verificar a relação entre as variáveis independentes x (preditoras) e variáveis dependentes y. Esta correlação representa o grau de impacto presente em cada variável x na variável analisada y. Recentemente, está sendo explorado o uso de variáveis intervalares para representar as propriedades do fenômeno. Esta relação ocorre devido ao aumento da informação presente no intervalo. Métodos que atuam com intervalos são necessários, uma vez que suas operações precisam comportar toda a representatividade fornecida pelos intervalos. Desta maneira, surgem métodos como Regressão Linear Intervalar (MARINO, 2002), (CARVALHO, 1995) e (NETO, 2010). Todos os trabalhos utilizam os conceitos de SIA (MOORE, 2003).

A aritmética intervalar multidimensional RDM-IA (do inglês, *Relative Distance Measure*)(PIEGAT,2017) surge para corrigir os problemas encontrados em SIA. Além de garantir as mesmas qualidades fornecidas pelo uso de intervalos. Um intervalo [x] em RDM-IA é representado da seguinte forma: $[x]=\{x : x=\underline{x}+\alpha_x(\overline{x}-\underline{x}), \alpha_x \in [0,1]\}$.

As vantagens caracterizadas pela RDM-IA é a inserção da variável multidimensional α que incorpora um novo parâmetro de incerteza do intervalo. Esta variável em que os valores ocultos estão presentes no intervalo [x] são acessíveis. As operações aritméticas definidas a partir da relação entre dois intervaos [x] e [y] são:

- **Adição:**

$$[x]+[y]=\{x+y : x+y=\underline{x}+\alpha_x(\overline{x}-\underline{x})+\underline{y}+\alpha_y(\overline{y}-\underline{y}), \alpha_x, \alpha_y \in [0,1]\} \quad (1)$$

- **Subtração:**

$$[x]-[y]=\{x-y : x-y=\underline{x}+\alpha_x(\overline{x}-\underline{x})-(\underline{y}+\alpha_y(\overline{y}-\underline{y})), \alpha_x, \alpha_y \in [0,1]\} \quad (2)$$

- **Multiplicação:**

$$[x]*[y]=\{x-y : x-y=(\underline{x}+\alpha_x(\overline{x}-\underline{x}))*(\underline{y}+\alpha_y(\overline{y}-\underline{y})), \alpha_x, \alpha_y \in [0,1]\} \quad (3)$$

-

- **Divisão**

$$[x]/[y]=\{x-y : x-y=(\underline{x}+\alpha_x(\overline{x}-\underline{x}))/(\underline{y}+\alpha_y(\overline{y}-\underline{y})), \alpha_x, \alpha_y \in [0,1]\} \quad (4)$$

Desta maneira, para [x] * [y], as operações básicas * ϵ $\{+,-,.,/\}$, em que a operação de divisão deve respeitar a restrição 0 ϵ [y].

Este trabalho desenvolve o método de CLRM-RDM (*Classical Linear Regression Model - Relative Distance Measure*), no qual introduz os conceitos de intervalos RDM para resolução de análises de regressão. O trabalho está organizado da seguinte forma: Na Seção 2 será apresentado o método de Regressão Linear. Na seção 3 apresentará a construção do CLRM-RDM, utilizando os conceitos de RDM conjuntamente com os da RL. Os resultados serão demonstrados a partir de exemplos apresentados no estado da arte, a fim de obter uma referência de avaliação. E por último a conclusão deste trabalho.

2. METODOLOGIA

Um modelo de regressão linear é apresentado na equação 5, sendo a variável Y representa a variável dependente, X a variável independente ou preditor desta regressão, β consistem nos estimadores desconhecidos do modelo e por último o ε representa o resíduo ou erro presente no modelo de regressão.

$$Y_i=\beta X_i+\varepsilon \quad (5)$$

Em que Y_i, X_i e β são representados através de matrizes. O método de RL fornece a menor solução possível através da soma do quadrado dos desvios da observação y_i (RAWLING, 2001) Desta maneira os parâmetros devem estimar o modelo de regressão de forma a minimizar o erro. A fim de estimar os parâmetros através de $\hat{\beta}$, apresenta-se a equação 6, no qual é construído em base nas variáveis independentes X_i existentes.

$$\hat{\beta}=(X^T X)^{-1} X^T Y \quad (6)$$

Sendo X_i uma matriz, então X^T e X^{-1} são a matriz transposta e inversa respectivamente. Por fim, a variável dependente estimada a partir dos parâmetros $\hat{\beta}$ e das variáveis independentes X_i é dada pela equação 7.

$$\hat{Y}_i=\hat{\beta} X_i \quad (7)$$

Todas as operações citadas acima são necessárias para realizar a correta adequação para utilizar-se de variáveis em RDM-IA. Conforme descrita na equação (PIEGAT,2017) a definição de um intervalo RDM e dado por [x]. Desta maneira utilizando variáveis independentes como valores em RDM, temos a representação $[X]$, desta maneira realizando os ajustes necessários nas equações 5,6 e 7, teremos as equações 8,9 e 10, respectivamente.

$$Y_i=\beta[X_i]+\varepsilon_i \quad (8)$$

$$\hat{\beta}=([X]^T X)^{-1}([X]^T Y) \quad (9)$$

$$[\hat{Y}]=[\hat{\beta}][X] \quad (10)$$

X^T representa a matriz transposta de X. A operação X^T deve conter uma propriedade para ser realizada de forma correta, sendo esta condição conhecida como **forte regularidade**. Diversos estudos apresentam a superestimação dos intervalos encapsuladores em SIA. Porém o mesmo problema de superestimação não ocorre em RDM-IA, uma vez que seus intervalos não são maiores do que o necessário.

Em que $[\hat{Y}]=\{\hat{Y}_0,\ldots,\hat{Y}_n\}$ e $\hat{Y}_i=\underline{\hat{y}_i}+\alpha_{\hat{y}_i}(\overline{\hat{y}_i}-\underline{\hat{y}_i})$ a representação em RDM da estimativa da variável observada. Note que todo o processo foi guiado a gerar números intervalares RDM, para assim poder usufruir das vantagens sistemáticas providas pela RDM-IA. As incertezas presentes em SIA, principalmente as relacio-

nadas a resolução de equações são completamente contornadas pela abordagem de Piegat(2017).

O uso de regressões com dados intervalares consistem em uma pequena parcela de problemas existentes dentro do escopo de problemas lineares. Este trabalho então buscou atender não somente a demanda de resolução de regressões lineares com dados intervalares, mas também modelar as mesmas resoluções para sistemas com dados compostos (real e intervalar) para assim poder manter a congruência com os problemas atualmente enfrentados.

3. RESULTADOS

A fim de validar o modelo de regressão linear com variáveis em RDM, será validado através de métricas que avaliam a qualidade do modelo matemático gerado pelo método CLRM-RDM. Os dados escolhidos para o experimento são os apresentados por BILLARD (2000), no qual consistem no registro da taxa de pulsação (Y), pressão sanguínea diastólica (X_1) e sistólica (X_2). Os dados estão disponíveis em BILLARD (2000). Será realizado um teste de Regressão Linear simples sobre a variável independente X_2. Serão comparadas com os resultados alcançados por BILLARD (2000) e NETO (2010).

As métricas de avaliação utilizadas são RMSE (*Root Mean Square Error*), Coeficiente de Correlação (r^2) e Coeficiente de Determinação (R^2). E todas são aplicadas diretamente sobre os valores estimados $\hat{Y}$.

Para o conjunto de dados utilizados no modelo de regressão estimado pelo método de CLRM-RDM é apresentado pela equação 11, que diferentemente das demais propostas, consiste em um único modelo de regressão para os dados intervalares. A Tabela 1 apresenta as métricas avaliativas para comparação com os demais métodos encontrados na literatura que tratam de resolução de problemas lineares, utilizando-se intervalos.

$$Y = [29.367708, 29.367708] + [0.38491412, 0.38491412] * [X_2] \quad (11)$$

Apesar do método de CLRM-RDM apresentar uma forma encapsulada para geração de modelos intervalares para análise de regressão, os resultados alcançados pelo trabalho de Billard (2000) e Neto (2010) são mais próximos da variável observada. Porém, sua forma de realização do cálculo desconsidera a informação correspondente no intervalo, uma vez que as operações não são realizadas da forma intervalar. CLRM-RDM apesar de conter um erro maior conforme apresentado na Tabela 1, demonstra a quantidade de erro numérico gerado na resolução do problema.

Tabela 1: Comparação dos métodos de resolução de Regressão Linear com variáveis intervalares

	$RMSE_{lower}(\%)$	$RMSE_{upper}(\%)$	$r^2_{lower}(\%)$	$r^2_{upper}(\%)$	$R^2(\%)$
Billard-LR	4.63	0.93	4.0	37.30	80.39
CCRM-LR	9.73	9.71	40.96	59.48	58.00
CLRM-RDM	14.22	14.22	3.10	20.30	64.39

Notavelmente, conforme Tabela 1, os valores de RMSE para o método de Billard são diferentes, pois sua representação intervalar não passa de uma representação, sem utilizar de fato as operações que desta estão definidas. Porém o modelo de treinamento gerado por Billard (2000) e Neto (2010) é mais exato para representar o conjunto de dados observados, conforme pode ser visto pelos indicadores apresentados na Tabela 1. Esta relação pode ocorrer, uma vez que ajustes intermediários em ambos os métodos são utilizados, visando deixar o intervalo

com diâmetro mínimo. No método CLRM-RDM desenvolvido, é somente utilizado a aritmética intervalar RDM em sua forma original.

4. CONCLUSÕES

O presente trabalho buscou apresentar a e desenvolver um método alternativo de análise de regressão linear com variáveis intervalares, visando utilizar os conceitos da aritmética intervalar RDM para contornar os problemas existentes na aritmética de Moore. Os métodos atualmente existentes que tratam o uso de variáveis intervalares no conjunto de dados, não utilizam corretamente as operações intervalares, inclusive realizam o cálculo dos limites do intervalo de forma independente.

O método criado neste trabalho denominado de CLRM-RDM manteve a integridade das operações definidas por Piegat (2017), como também a íntegridade de todas as métricas realizadas sobre análises de regressão. Apesar dessa certificação das operações, os resultados não foram significativos, contendo maior erro numérico e criando um modelo matemático que não representa corretamente o conjunto de dados analisados. Como trabalhos futuros, será refinado o método e utilizar das operações de ajustes utilizados nos trabalhos encontrados na literatura.

5. REFERÊNCIAS BIBLIOGRÁFICAS

BILLARD L., DIDAY E., "**Regression analysis for interval-valued data**," in Data Analysis, Classification, and Related Methods, pp. 369–374, Springer, 2000

CARVALHO F., "**Histograms in symbolic data analysis,**" Annals of Operations Research, vol. 55, no. 2, pp. 299–322, 1995

GOLDBERG D., "**What every computer scientist should know about floating-point arithmetic**" ACM Computing Surveys (CSUR), vol. 23, no. 1, pp. 5–48, 1991.

MARINO M., PALUMBO F., "**Interval arithmetic for the evaluation of imprecise data effects in least squares linear regression,**" Statistica Applicata, vol. 14, no. 3, pp. 277–291, 2002.

MOORE R., LODWICK. W., "**Interval analysis and fuzzy set theory,**" Fuzzy sets and systems, vol. 135, no. 1, pp. 5–9, 2003.

NETO E., CARVALHO F., "**Constrained linear regression models for symbolic interval-valued variables**," Computational Statistics & Data Analysis, vol. 54, no. 2, pp. 333–347, 2010.

RATSCHEK H., ROKNE J., "**New computer methods for global optimization.**" Dosseldorf, Alemanha: Horwood Chichester, 1988.

RAWLINGS J., PANTULA S., e DICKEY D., "**Applied regression analysis: a research tool**". Springer Science & Business Media, 2001.

PIEGAT A., and LANDOWSKI M., "**Is an interval the right result of arithmetic operations on intervals?,**" International Journal of Applied Mathematics and Computer Science, vol. 27, no. 3, pp. 575–590, 2017