ANÁLISE DE MÉTODOS ESTATÍSTICOS PARA CONJUNTO DE DADOS EM GRANDE ESCALA – BIG DATA

SILVA, Vanessa Gorete da[1]; COSTA, Danielle[2]; COSTA, Luzia Aparecida da[3]

[1]Estudante do curso de Engenharia Elétrica do Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais (IFMG) -*Campus* Formiga, voluntário (PIBIC). E-mail: vanessa_g_silva@yahoo.com
[2]Professora orientadora do Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais (IFMG) - *Campus* Formiga. E-mail: danielle.costa@ifmg.edu.br
[3] Professora orientadora do Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais (IFMG) - *Campus* Formiga. E-mail: luzia.costa@ifmg.edu.br

Resumo: Com os avanços das tecnologias na última década, a quantidade de dados gerados e registrados cresceu muito em praticamente todos os setores da indústria e da ciência. Tal fato oferece oportunidades sem precedentes para a tomada de decisões baseada em dados e a descoberta de conhecimento. No entanto, a tarefa de fazer essa análise em grande escala representa desafios significativos e exige métodos estatísticos inovadores projetados especificamente para maior rapidez e maior eficiência. O presente projeto de pesquisa tem como objetivo investigar os métodos estatísticos atualmente disponíveis para grande volume de dados – *Big Data*. Trata-se de um trabalho que envolve pesquisa bibliográfica, o qual espera-se como principal resultado, apresentar uma visão geral dos métodos estáticos usados para um conjunto de dados em grande escala.

Palavras-chave: Big Data. Métodos estatísticos. Regressão Linear.

1 INTRODUÇÃO

O rápido desenvolvimento de tecnologias na última década permitiu aos pesquisadores gerar e coletar dados com tamanhos e complexidades sem precedentes em todos os campos da ciência e engenharia, da academia à indústria, apresentando desafios significativos na descoberta de conhecimento.

O atributo volume é a característica mais significativa no conceito de *Big Data* que faz referência à dimensão do volume de dados que, se utilizado habilmente, torna-se relevante à tradução de informações em vantagens comerciais, por exemplo (MCAFEE; BRYNJOLFSSON, 2012).

A análise de *Big Data* incorpora o processo de examinar conjuntos massivos de dados para descobrir padrões escondidos, correlações desconhecidas, além de algoritmos eficazes para análises de populações inteiras de dados e não mais amostras definidas convencionalmente.

Big Data são especialmente desafiadores porque alguns deles não foram coletados para abordar uma questão específica. A disciplina científica da estatística traz técnicas e modelos sofisticados para lidar com essa questão. Segundo Triola (1999), a estatística é uma

coleção de métodos para planejar experimentos, obter dados e organizá-los, resumi-los, analisá-los, interpretá-los e deles extrair conclusões. Sendo assim, o objetivo deste trabalho é fornecer um estudo sobre os métodos estatísticos adequados à análise de *Big Data*, buscando ampliar as formulações teóricas sobre o tema e respondendo à questão de quais métodos podem ser aplicáveis às análises de forma a gerar conhecimento útil.

Espera-se, com isso, contribuir para o estado da arte produzindo um referencial teórico que possa ser utilizado, por alunos e pesquisadores, como subsídio para novas pesquisas e desenvolvimentos na área.

Além dessa primeira seção, este artigo está organizado em outras três da seguinte forma: na seção 2, é apresentada a metodologia aplicada no projeto de pesquisa; na seção 3, são descritos os resultados e discussão e, finalmente, na seção 4, são apresentadas as conclusões do trabalho.

2 MATERIAIS E MÉTODOS

A pesquisa é bibliográfica desenvolvida a partir de materiais publicados em livros, artigos, dissertações e teses, constituindo o procedimento básico para os estudos em *Big Data* e Estatística, pelos quais se busca o domínio do estado da arte sobre esses temas.

Caracteriza-se como um estudo descritivo-exploratório: descritiva, pois evidencia a descrição dos conceitos que permeiam a temática e exploratória, pois busca identificar o que está sendo utilizado em estatística para a análise de dados em grande escala.

3 RESULTADOS E DISCUSSÃO

Tratando-se de abordagens que envolvem a estatística aplicável a *Big Data*, pode-se classificá-las quanto:

- à divisão e método de conquista, que se refere à divisão de um extenso conjunto de dados em blocos menores, gerenciáveis computacionalmente, aos quais se aplica uma estratégia estatística apropriada a cada bloco e também para combinar seus resultados (HOROWITZ, 2001);
- a metodologia como método fino a grosso, que trata da criação de algoritmos para trabalhar com grandes escalas através do arredondamento de parâmetros. O método fino a grosso consiste em colettar informações e, baseando-se em

técnicas de análise de variância, estudar o comportamento da homocedasticidade das miniamostras (HELWIG; MA, 2016);

- ao método de amostragem, no qual se retira uma subamostra do conjunto de dados original, utilizando-a com objetivo de examinar estimativas do modelo proposto (CHEN; XIE, 2014);
- aos modelos lineares, devido à sua ampla abrangência em diferentes áreas, bem como sua utilidade no processo de planejamento de pesquisas e análises de resultados, traduzindo-se, portanto, em uma metodologia aplicável a *Big Data* (RENCHER; SCHAALJE, 2008). Após investigações e partindo de tal assertiva, nota-se o caráter abrangente da Regressão Linear para análise em *Big Data*, haja vista sua compatibilidade de trabalho em variáveis multivariadas.

O Método de Regressão Linear apresenta-se em duas formas: Regressão Linear Simples e a Regressão Linear Múltipla. Nos casos em que ocorrem correlação entre variáveis, é possível determinar o comportamento dessas em funções de outras, elaborando um artifício eficaz para análises de pequenos e grandes volumes de dados (FIELD, 2009).

Na Regressão Linear Simples, é possível realizar uma análise com duas variáveis, relacionando uma variável dependente a uma variável independente, e dessas extrair informações (RENCHER; SCHAALJE, 2008). A equação que traduz tal modelo é explicitada a seguir:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (1)$$

Na Equação (1), o Y_i é a variável dependente, X_i a variável independente, β são os parâmetros a serem estimados e ε_i representa fatores residuais e possíveis erros de medição, não passíveis de serem linearmente explicados. Entretanto, tratando-se de análises em *Big Data* o Modelo de Regressão Linear Simples não corresponde às demandas, pois em tal contexto é comum aparecerem *n* variáveis e esse modelo lida com estimações bivariadas.

Na Regressão Linear Múltipla, obtém-se a presença de uma variável resposta *y*, relacionada linearmente com diversas variáveis preditoras, objetivando melhorar predições (COSTA, 2015). A equação que traduz tal modelo é:

$$Y = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{i\,p} + \varepsilon_i \quad (2)$$

A Equação (2) refere-se ao Modelo de Regressão Linear Múltipla, em que Y representa a variável resposta, X_i são valores de i-enésima observação das *p* variáveis independentes, βp são parâmetros e ε_i representa erros aleatórios.

Uma vez feito o ajuste do modelo através dos dados de *Big Data*, uma das formas de avaliar a qualidade desse ajuste é através do coeficiente de determinação R^2, obtido entre uma razão da variação do modelo ajustado e a variação total. Como é uma razão, esse coeficiente está no intervalo [0,1] e, basicamente, tal coeficiente indica quanto o modelo foi capaz de explicar os dados em estudo.

Portanto, quando se trata de um conjunto de dados grandes, o Modelo de Regressão Linear Múltipla apresenta-se mais apropriado, pois a variável resposta frequentemente é influenciada pela presença de inúmeras variáveis preditoras. Dado tal fato, é preciso selecionar qual o critério que melhor se adequa à análise e, a partir deste ponto, aplicar o critério escolhido a fim de extrair dos dados dispostos as mais excelentes informações. Ressalta-se que neste trabalho estão sendo examinados os melhores critérios a serem adotados.

4 CONCLUSÕES

O estudo realizado evidencia que o uso da Regressão Linear é de extrema importância como ferramenta estatística para análise de conjunto de dados grande, pois a seleção de variáveis é um meio para se chegar a um modelo matemático. Destaca-se a importância de se obter um modelo mais adequado para alcançar predições ou que explique de forma correta a relação entre as variáveis em questão. Quando existem diversas variáveis ou um volume grande de informações, como é caso das análises em *Big Data*, é sugerido usar métodos de seleção para eliminar as variáveis com efeitos insignificantes, e, então, o conjunto reduzido pode ser investigado novamente, sendo escolhido o modelo que apresentar melhor R^2. Logo, espera-se obter o melhor modelo possível, de modo que esteja fundamentado o suficiente para estimar e/ou prever futuras situações.

REFERÊNCIAS

COSTA, Luzia Aparecida da. **Novo estimador de cumeeira de Rao com aplicação em seleção genômica.** 2015. Tese (Doutorado em Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras, 2015.

CHEN, Xueying; XIE, Min-ge. **A split-and-conquer approach for analysis of extraordinarily large data.** Statistica Sinica, v.24, p.1655-1684, 2014.

FIELD, Andy. **Descobrindo a estatística usando o spss.** 2 ed. Porto Alegre: Artmed, 2009.

MCAFEE, A.; BRYNJOLFSSON, E. Big Data: the management revolution. **Harvard Business Review**, v. 90, n. 10, p. 60-68, 2012.

HELWIG, Nathaniel E.; MA, Ping. Smoothing spline ANOVA for super-large samples: scalable computation via rounding parameters. **Statistics And Its Interface**, v. 9, n. 4, p.433-444, 2016.

HOROWITZ, Joel L. The bootstrap. **Handbook Of Econometrics**, p.3159-3228, 2001.

RENCHER, C. A.; SCHAALJE, B. G. **Linear models in statistics**. 2 ed. New Jersey: J. Wiley Sons, 2008.

TRIOLA, Mario F. **Introdução à estatística.** 7 ed. Rio de Janeiro: Ltc, 1999.

Como citar este trabalho:
SILVA, V. G.; COSTA, D.; COSTA, L. A. Análise de métodos estatísticos para conjunto de dados em grande escala – *Big Data*. *In*: SEMINÁRIO DE PESQUISA E INOVAÇÃO (SemPI), III., 2019. Formiga. **Anais eletrônicos** [...]. Formiga: IFMG – *Campus* Formiga, 2019. ISSN – 2674-7111.