https://eventos.utfpr.edu.br//sicite/sicite2020

Cuidados ao lidar com modelos de locação: Uma perspectiva GAMLSS

Care when dealing with location models: A GAMLSS perspective

Renan Jacob de Carvalho
renanjacobcarvalho@hotmail.com
Universidade Tecnológica Federal do Paraná, Apucarana, Paraná, Brasil

Thiago G. Ramires
thiagogentil@gmail.com
Universidade Tecnológica Federal do Paraná, Apucarana, Paraná, Brasil

Ana Júlia Righetto
ajrighetto@gmail.com
Instituto Agronômico do Paraná, Apucarana, Paraná, Brasil

Luiz Ricardo Nakamura
luiz.nakamura@ufsc.br
Universidade Federal de Santa Catarina, Florianópolis, Santa Catarina, Brasil

RESUMO

Neste artigo, apresentamos uma discussão sobre modelos de regressão, especialmente aqueles pertencentes a classe de localização. A principal motivação desta pesquisa é mostrar que, mesmo usando simples distribuições que algumas vezes não são assim flexível, podemos obter melhores resultados quando comparados com distribuições complexas. Para melhorar os modelos, estamos adotando o aditivo generalizado modelos para localização, escala e estrutura de forma, o que permite vários parâmetros (características) dos modelos probabilísticos, como média, moda, variância e outros. Lidar com efeitos não lineares, funções suaves, tais como splines penalizadas, são incorporadas no estruturas de regressão. No final, as comparações são feitas usando três conjuntos de dados reais, mostrando que modelos probabilísticos simples e interpretáveis são preferíveis, ao usar um mais complexo estrutura de regressão, do que modelos probabilísticos complexos.

PALAVRAS-CHAVE: Regressão. Splines penalizadas. Modelos probabilísticos.

ABSTRACT

In this paper we present a discussion of regression models, especially those belonging to the location class. The main motivation of this survey is to show that, even using simple distributions that some times are not so exible, we can get better results when compared with complex distributions. To improve the models, we are adopting the generalized additive models for location, scale and shape framework, which allows to t several parameters (characteristics) of the probabilistic models, like mean, mode, variance and others. To deal with non-linear effects, smooth functions, such penalized splines, are incorporated in the regression structures. At end, comparisons are made by using three real data sets, showing that simple and interpretable probabilistic models are preferable, when using a more complex regression structure, than complex probabilistic models

KEYWORDS: Regression. Splines. Parsimony principle

Recebido: 19 ago. 2020.
Aprovado: 01 out. 2020.

INTRODUÇÃO

Muitos estudos são publicados com a missão de melhorar os ajustes e predições feitas por estudos prévios e como exemplos podemos citar aplicações de tais modelos a bases de dados de várias áreas como é o caso de (KORKMAZ et al., 2019a) que propõe um modelo de regressão Log-OLLMOGHN sobre 4 parâmetros para dados de HIV, (KORKMAZ et al., 2019b) também introduz LHj-W, um modelo de regressão sobre 5 parâmetros para dados sobre expectativa de vida de pacientes submetidos à transplantes do coração, já (CORDEIRO et al., 2018) apresenta o LXG-W sobre 3 parâmetros para modelar dados de ocorrência de Retinopatia em diabéticos, (YOUSOF et al. 2019) por sua vez apresenta o modelo de regressão TLGBXII com 5 parâmetros para modelar dados de tempo de falha do epóxi sólido em aplicações elétricas, dentre outros.

Os modelos apresentados anteriormente, todavia, possuem limitações em comum que pode ser corrigida a fim de proporcionar um melhor ajuste, os referidos trabalhos possuem apenas o termo de locação em sua estrutura de regressão, logo, carrega suas características como determinante único. Nesse sentido, esses modelos podem apresentar bons resultados mesmo reproduzindo o comportamento de somente um parâmetro e a explicação para isso decorre do fato que os demais estão relacionados com significantes medidas do modelo como média, variância, simetria e curtose, dessa forma, quando estimada a máxima verossimilhança (*maximum likelihood*) essas medidas citadas são também modeladas.

Para propor um modelo de maior precisão e mais fácil compreensão como é o caso deste estudo, que apresenta a aplicação de GAMLSS para essas bases de dados, os métodos de comparação utilizados como modos de verificação são AIC e BIC, os quais são estabelecidos pelo log-verossimilhança (l), $\boldsymbol{df}$ que indicam os degraus de liberdade do modelo, k é o degrau de penalidade, sendo expressos por $-\mathbf{2}.\boldsymbol{l}(\boldsymbol{\theta}) + \boldsymbol{k} \times \boldsymbol{df}$, para $\boldsymbol{k} = \mathbf{2}$ ou $\boldsymbol{k} = \mathbf{log}\,(\boldsymbol{n})$. AIC e BIC são basicamente *log-likelihood* adicionado de uma penalização.

Todas as informações em uma amostra são relevantes para inferências do valor dos parâmetros de um modelo, isso pode ser notado considerando um modelo normal em que o parâmetro de locação é independente da medida de variação e ao comparar isso com a classe de modelos de localização, observa-se que não são consideradas todas as configurações existentes para a construção do modelo, isso significa que esses modelos consideram apenas as informações do parâmetro de locação que pode ou não representar a média.

METODOLOGIA

A utilização do GAMLSS (*generalized additive models for location, scale and shape*) em vez das classes mais tradicionais como os modelos de Regressão por locação referidos anteriormente, pode estabelecer uma evolução nos modelos tendo em vista esses podem ser caracterizados como:

$$\boldsymbol{Y} = \boldsymbol{\mu}(\mathbf{v}) + \boldsymbol{Z} \qquad (1)$$

Em que $\boldsymbol{Y}$ é a variável resposta e depende do parâmetro de locação μ(v) que por sua vez, depende de $\boldsymbol{v}$, o vetor de variáveis explanatórias. O vetor $\boldsymbol{Z}$, que segue uma determinada distribuição, não depende diretamente do vetor das variáveis explanatórias (v).

Para uma determinada distribuição, como é o caso de Reverse Gumbel (RG), composta de dois parâmetros, o parâmetro de locação $-\infty < \boldsymbol{\mu} < +\infty$ e o parâmetro de escala $\mathbf{0} < \boldsymbol{\sigma} < +\infty$, a variável resposta segue $\boldsymbol{Y} \sim \boldsymbol{RG}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ uma vez que função pdf para essa distribuição é dada por:

$$\boldsymbol{f}(\boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \frac{\mathbf{1}}{\boldsymbol{\sigma}} \mathbf{exp}[-\boldsymbol{z} - \boldsymbol{exp}(-\boldsymbol{z})] \quad \text{para} \quad -\infty < \boldsymbol{y} < +\infty \tag{2}$$

Em que $\boldsymbol{z} = (\boldsymbol{y} - \boldsymbol{\mu})/\boldsymbol{\sigma}$. Considerando que na eq.(1) a variável $\boldsymbol{Z}$ segue a distribuição Gumbel reverso $\boldsymbol{RG}(\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\sigma})$, consequentemente, a variável resposta $\boldsymbol{Y}$ também seguirá a distribuição $\boldsymbol{\theta}(\boldsymbol{\mu}(\mathbf{v}), \boldsymbol{\sigma})$ o que indica que a moda está sendo modelada diretamente enquanto a média e mediana são modeladas indiretamente.

Por outro lado, modelando somente o termo de locação, pode afetar a confiabilidade do resultado e aumentar o erro padrão. A fim de evitar tais dúvidas, uma maneira mais viável de se contornar essa dificuldade seria modelar os parâmetros de simetria, variância entre outros parâmetros relacionados, também com as variáveis explanatórias.

Nesse caso, reescrevendo o modelo de distribuição Gumbel reverso introduzindo a classe de GAMLSS, resulta em:

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\sigma} \end{bmatrix} = \begin{bmatrix} \boldsymbol{g}(\boldsymbol{X_1 B_1}) \\ \boldsymbol{\sigma} \end{bmatrix} \tag{3}$$

Em que $g(.)$ é a função de ligação, B_1é o vetor de variáveis explanatórias e X_1 é uma matriz conhecida de ordem $n \times (m_1 + 1)$, onde m_1 é o número de variáveis explanatórias. Generalizando a estrutura de modo a estabelecer a relação entre as variáveis explanatórias e os parâmetros do modelo usando de funções de suavização como *p-splines, cubic splines, ridge-lasso* etc. a estrutura de regressão pode ser determinada por:

$$g_r(\theta_r) = X_r \mathrm{B}_r + \sum_{j=1}^{J_r} s_{jr}(x_{jr}) \tag{4}$$

Na eq.(4) s_{jr}, representa a função de suavização das variáveis explanatórias que podem ser mais detalhadas por (EILERS; MARX, 1996).

RESULTADOS E DISCUSSÕES

Para a construção de estruturas de regressão usando a classe GAMLSS, um clássico banco de dados usado pela literatura servirá de motivação, considerando a regressão baseada na distribuição reversa de Gumbel. Para isso, "*Voltage*" é um conjunto de observações propostas por (LAWLESS, 2003) sobre o tempo de vida do epóxi sólido em aplicações elétricas. O banco de dados, basicamente, expõe o

tempo de falha do material (min), considerando três níveis de tensão elétrica (x_i) e por meio de $n = 60$ observações, nas quais, 6 delas foram classificadas como censuradas.

Para fins de comparação das melhores aproximações do modelo, será utilizada a distribuição $Y = \log(tempo) \sim RG(\mu, \sigma)$, assim, a principal análise a ser feita é como X impacta no tempo de falha, considerando esse como uma variável contínua. Analisando as densidades através da fig(1), é possível perceber uma não linearidade entre μ e X, tendo em vista que a moda para $x_i = 57{,}5$ e $x_i = 55{,}0$ possuem certa semelhança entre si, mas diferem da moda de $x_i = 52{,}5$.

Figura 1 – Densidades para cada nível de tensão

Fonte: Autoria Própria (2019).

Depois de feita essa consideração, é possível estabelecer um modelo pela seleção de variáveis a fim de chegar ao modelo final que represente Y na eq.(5). Além disso, a tabela [1] mostra a comparação dos valores de AIC e BIC dos modelos para esse banco de dados, demonstrando que considerando uma distribuição simples é possível obter bons modelos.

$$\mu_i = 15.646 + pb(x_i) \quad e \quad \log(\sigma_i) = 5.38 + pb(x_i) \tag{5}$$

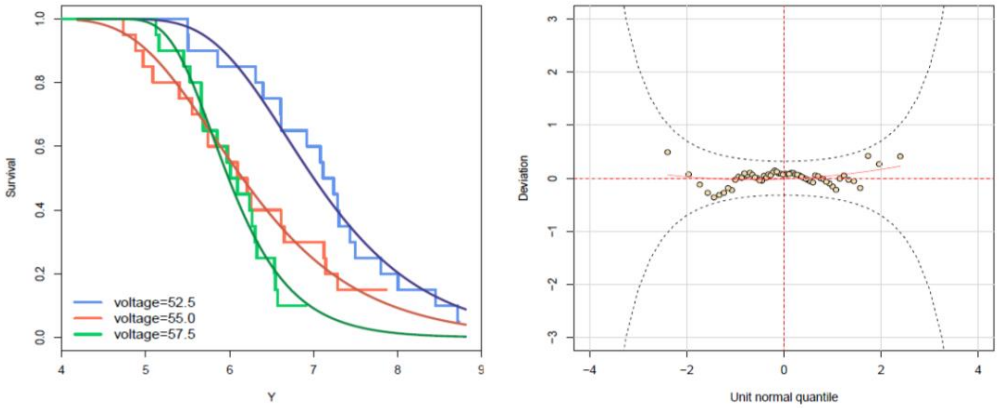
Tabela 1 – Valores de AIC e BIC correspondentes a cada modelo

Modelo	AIC	BIC
RG (gamlss)	157.6	168.6
LOLLGHN	166.4	174.8
Log-Weibull (Gumbel)	173.4	179.7
LGE-W	168.6	179.1
LZBOLL-GHN	166.2	176.7
LWMOW	173.5	184.0
LKwGR	177.4	187.8
LOLLFr	164.3	172.7

Fonte: Autoria própria (2019).

A figura 2(a) mostra a curva de sobrevivência do modelo adotado e a figura 2(b) a análise residual baseada em WP. Ambas se complementam e demonstram a qualidade do modelo adotado

Figura 2- (a) curvas de sobrevivência da distribuição RG e (b) o WP



Fonte: Autoria própria (2019).

CONCLUSÃO

Apresentados os modelos de regressão, foi feita uma revisão do comportamento dos dados a partir da utilização de estruturas lineares e não lineares bem como a utilização de todos os parâmetros. analisando os resultados foi possível perceber que o uso de modelos mais simples com parâmetros interpretávies obteve melhor resposta quando comparados com modelos mais complexos. conclui-se que, com o conhecimento da classe gamlss, grande parte dos modelos mais complexos (com muitos parâmetros) não é necessária.

AGRADECIMENTOS

Agradeço especialmente ao apoio da Fundação Araucária, da qual fui auxiliado financeiramente durante o período de Iniciação científica. Também ao orientador prof. dr. Thiago G. Ramires por dedicar seu tempo ao guiar nos estudos a fim de gerar trabalhos acadêmicos de excelência. Por fim, à Universidade Tecnológica Federal do Paraná.

REFERÊNCIAS

CORDEIRO, G. M.; BOURGUIGNON, M.; ORTEGA, E. M.; RAMIRES, T. G. **General mathematical properties, regression and applications of the log-gamma-generated family**. *In:* Communications in Statistics-Theory and Methods, 2018, 47, p. 1050-1070.

EILERS, P.H.; MARX, B.D. **Flexible smoothing with B-splines and penalties**. *In:* Statistical Science, 1996, p. 89-121.

KORKMAZ, M. C. ; ALTUN, E. ; ALIZADEH, M. ; YOUSOF, H. M. **A new flexible lifetime model with log-location regression modeling, properties and applications**. *In:* Journal of Statistics and Management Systems, 2019, p. 871-891.

KORKMAZ, M. C.; ALTUN, E.; YOUSOF, H. M.; HAMEDANI, G. G. **The odd power Lindley generator of probability distributions: properties, characterizations and regression modeling**. *In:* International Journal of Statistics and Probability, 2019, 8, p. 70-89.

LAWLESS, J. F. **Statistical models and methods for lifetime data**. New York: Wiley, 2003

YOUSOF, H. M.; RASEKHI, M.; ALTUN, E.; ALIZADEH, M. **The extended odd Frechet family of distributions**: properties, applications and regression modeling. *In:* Int J Math Comput, 2019, p. 1-16.