MLG duplo

Deiverson Eduardo Oliveira de Almeida e Lucas Avila Moreira de Paula

16/01/2023

Contents

1 Abstract

Nesse trabalho nos propomos a resolver problemas de modelagem para variáveis explicativas que não necessariamente seguem distribuição normal e possuem variância inconstante. Nesse sentido, já haviam métodos que faziam transformações para resolver esse problema (como a box-cox), mas nem sempre eram eficazes. Através de propriedades da família exponencial e de funções de ligação, iremos modelar, simultaneamente, a média e a variância dos dados. Ao final, para mostrar a eficácia da tecnica, iremos simular um conjunto de dados heterocedástico e comparar a precisão dos modelos lineares generalizados com os modelos generalizados duplos aqui propostos.

2 Introdução

Modelos lineares generalizados são uma ótima forma de modelarmos um conjunto de dado, mesmo quando eles não seguem distribuição normal. Contudo, há neles uma limitação. Partimos do pré-suposto de que se verifica homocedasticidade, isto é, que a variância é constante. Quando tal suposto não é válido, pode ser útil utilizar os Modelos Lineares Generalizados Duplos, proposto por Nelder e Wedderburn em 1972, no qual fazemos a modelagem da média e da variância simultaneamente

3 Especificação do modelo

Seja $Y_1, Y_2, ..., Y_n$ variáveis aleátorias pertecentes à família exponencial da forma

$$f(y_i|\theta_i, \phi_i) = exp\{\phi[y_i\theta_i - b(\theta_i)] + c(y_i, \phi_i)\}$$

e a função de suavização podendo ser escrita como

$$c(y_i, \phi_i) = d(\phi_i) + \phi_i a(y_i) + u(y_i)$$

Assim, a função log-verossimilhança será

$$L(\theta) = \sum_{i=1}^{n} \{\phi_i t_i + d(\phi_i) + u(y_i)\}$$

em que

$$t_i = y_i\theta_i - b(\theta_i) + a(y_i)$$

Se fixarmos θ a expressão acima coincide com um modelo da famiília exponencial com respostas independentes $T_1, T_2, ..., T_n$, parâmetros canônicos $\phi_1, ..., \phi_n$ e parâmetro de dispersão fixo igual a 1.

Aplicando as propriedades da família exponencial, temos que:

$$\mu_{T_i} = E(T_i) = -d''(\phi_i) e Var(T_i) = -d''(\phi_i)$$

Na tabela abaixo podemos ver esses valores para as famílias Normal, Normal inversa e Gama.

4 Estimação

Para o parâmetro β, temos a função escore abaixo:

$$\mathbf{U}_\beta = \mathbf{X}^T \mathbf{\Phi} \mathbf{W}^{1/2} \mathbf{V}^{-1/2} (\mathbf{y} - \boldsymbol{\mu});$$

Derivação de algumas quantidades para distribuições da família exponencial.

	Normal	Normal inversa	Gama
t_i	$y_i\mu_i - \frac{1}{2}(\mu_i^2 + y_i^2)$	$-\{y_i/2\mu_i^2 + \mu_i^{-1} + (2y_i)^{-1}\}$	$\log(y_i/\mu_i) - y_i/\mu_i$
$d(\phi)$	$\frac{1}{2}\log\phi$	$\frac{1}{2}\log\phi$	$\phi\log\phi - \log\Gamma(\phi)$
$d'(\phi)$	$(2\phi)^{-1}$	$(2\phi)^{-1}$	$(1 + \log\phi) - \psi(\phi)$
$d''(\phi)$	-$(2\phi^2)^{-1}$	-$(2\phi^2)^{-1}$	$\phi^{-1} - \psi'(\phi)$

Figure 1: Tabela retirada do livro "Modelos de Regressão com apoio computacional" de Gilberto A. Paula, página 163

e a matriz de informação de Fisher:

$$\mathbf{K}_{\beta\beta} = \mathbf{X}^T\mathbf{\Phi}\mathbf{W}\mathbf{X}.$$

Para obtermos a função escore do parâmetro γ, calculamos a seguinte derivada:

$$\frac{\partial L(\boldsymbol{\theta})}{\partial\gamma_j} = \sum_{i=1}^{n}\frac{z_{ij}}{h'(\phi_i)}\{t_i + d'(\phi)\}.$$

Em forma matricial, temos

$$\mathbf{U}_\gamma = \mathbf{Z}^T\mathbf{H}_\gamma^{-1}(\mathbf{t} - \boldsymbol{\mu}_T).$$

Já a matriz de informação para γ pode ser obtida calculando a derivada abaixo:

$$\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\gamma_j\partial\gamma_l} = -\sum_{i=1}^{n}\frac{z_{ij}z_{il}}{\{h'(\phi_i)\}^2}\left[d''(\phi_i) - \frac{h''(\phi_i)}{h'(\phi_i)}\ \{t_i + d'(\phi)\}\right].$$

O resultado acima pode ser escrito de forma matricial:

$$\mathbf{K}_{\gamma\gamma} = \mathbf{Z}^T\mathbf{V}_\gamma\mathbf{H}_\gamma^{-2}\mathbf{Z}.$$

Sendo assim, a matriz de informação de Fisher para $\boldsymbol{\theta}$ é:

$$\mathbf{K}_{\boldsymbol{\theta}\boldsymbol{\theta}} = \mathrm{diag}\{\mathbf{K}_{\beta\beta}, \mathbf{K}_{\gamma\gamma}\}.$$

Sob as condições de regularidade, temos que

$$\hat{\beta} \sim \mathrm{N}_p(\beta, \mathbf{K}_{\beta\beta}^{-1}) \quad \text{e} \quad \hat{\gamma} \sim \mathrm{N}_q(\gamma, \mathbf{K}_{\gamma\gamma}^{-1}),$$

desde que n seja suficientemente grande. Além disso, $\hat{\beta}$ e $\hat{\gamma}$ são assintoticamente independentes.

5 Aplicação em dados simulados

Geramos uma amostra de 800 variáveis aleatórias normais com média $\{1, 2, ..., n\}$, e variâncias $1^{1.5}, 2^{1.5}, ..., 3^{1.5}$.

Chamaremos essa variância de σ^2 e o vetor de médias de N

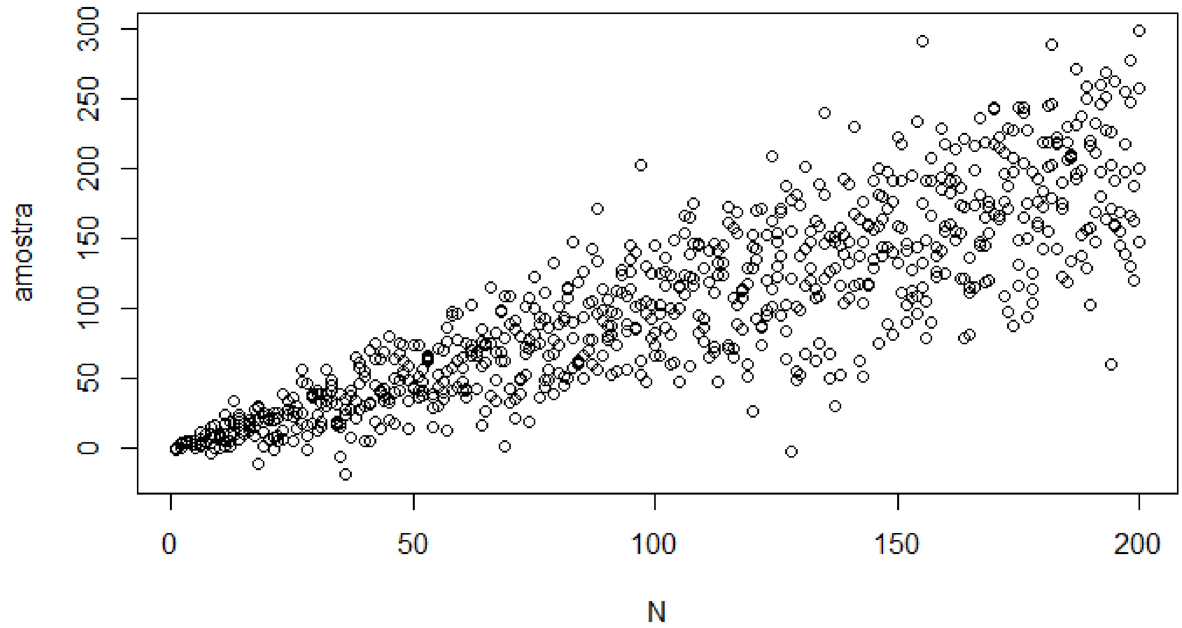

Figure 2: Gráfico mostrando que a série claramente não é homocedástica

Então aplicamos um Modelo Linear Generalizado, com família Normal e ligação identidade, resultando em:

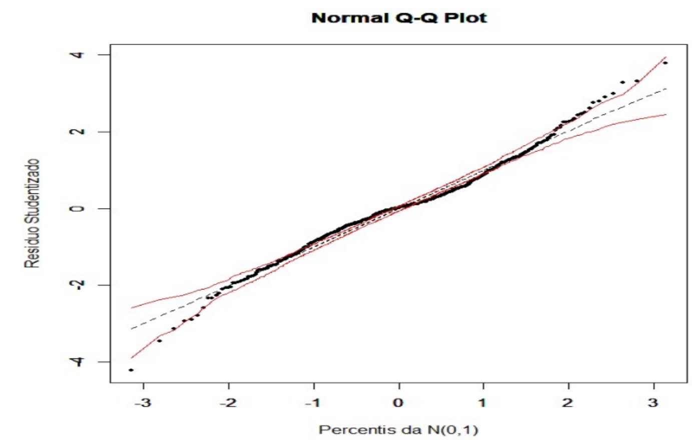

Figure 3: Envelope com muitos pontos foras, claramente não é adequado

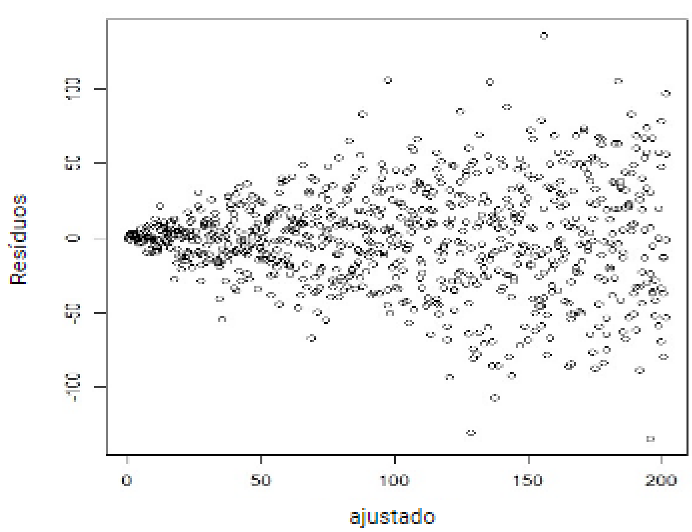

Figure 4: Resíuos do MLG contra seus valores ajustados. Claramente há heterocedasticidade

Mas se usarmos MLG duplo e escolhendo para modelagem da variância e média o vetor N com intercepto, teremos os seguintes resultados

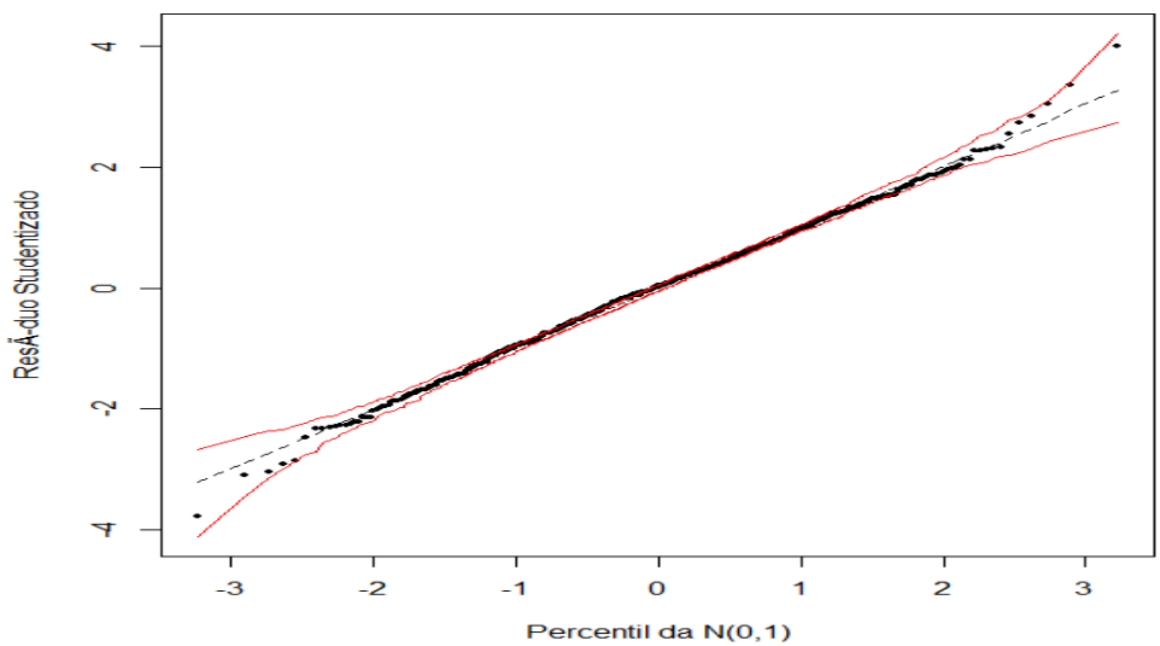

Figure 5: Envelope do valor esperado abriga praticamente todos os pontos, como era desejado

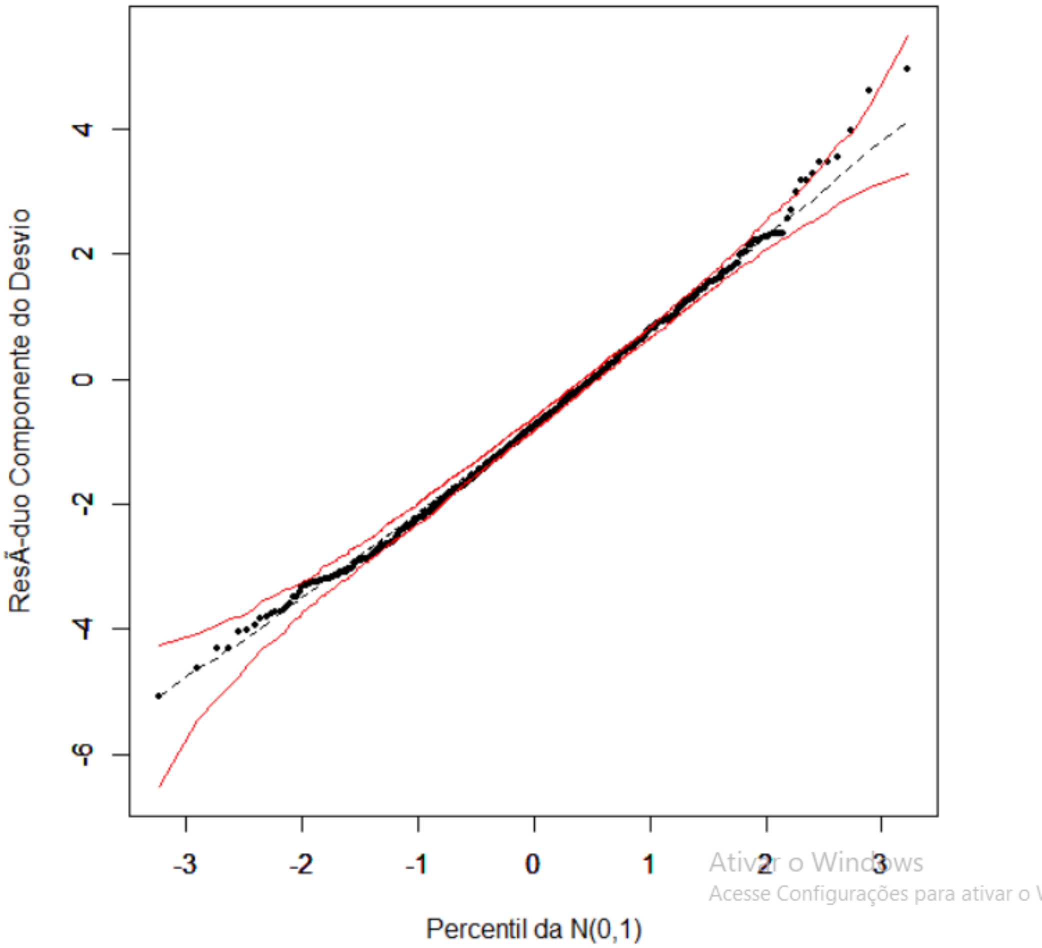

Figure 6: Envelope para dispersão. Como esperado, praticamente todos os valores estão dentro do envelope

6 Conclusão

Verificamos que os Modelos Lineares Generalizados Duplos são de fato muito úteis. A diferença na qualidade do ajuste para os dados foi inegável. Quando o pressuposto da homocedasticidade foi quebrado o Modelo Linear Generalizado desempenhou muito mal, como era esperado. Por outro lado, o modelo Modelo Linear Generalizado Duplo funcionou perfeitamente, tendo ótimos envelopes tanto para a modelagem da média quanto da dispersão

7 Referências

PAULA, Gilberto Alvarenga. Modelos de regressão: com apoio computacional.
SMYTH, GORDON K. Generalized Linear Models with Varying Dispersion
CAVALARO, Lucas Leite Um procedimento para seleção de variáveis em modelos lineares generalizados duplos
MCCULLAGH, P. and Nelder, J.A. 1989. Generalized Linear Models
SILVA, Paulo Henrique Dourado. Double Generalized Linear Models using SAS®: The %doubleglm macro
BORBA, Marcus Vinicius Teixeira. Modelos lineares generalizados duplos e aplicações
Paula, Gilberto Alvarenga. On diagnostics in double generalized linear models