Revista da Estatística UFOP, Vol III(3), 2014, ISSN 2237-8111
Edição especial: 59ª Reunião Anual da Regional Brasileira da Sociedade Internacional de Biometria - RBRAS
Trabalhos apresentados na versão poster.

Modelo Linear Generalizado Exponencial Potência

Cristian Villegas [1] [2]

1 Introdução

Os modelos lineares normais são amplamente aplicados em diversas áreas do conhecimento para modelar a média de dados contínuos que possuem uma distribuição simétrica. No entanto, quando temos presença de dados atípicos a distribuição exponencial potência surge como uma alternativa interessante à distribuição normal, já que possui caudas mais pesadas e mais curtas do que a distribuição normal e portanto ajusta melhor os dados atípicos. Com base em modelos lineares generalizados (MLG) [6] e modelos lineares simétricos (MLS) [5] estudamos o modelo linear generalizado exponencial potência (MLGEP). A ideia principal desses modelos é modelar dados contínuos com distribuição exponencial potência com alguma função de ligação entre a média e um conjuntos de preditores lineares. Em particular, comparamos o modelo linear normal e modelo linear exponencial potência para diferentes funções de ligação a um conjunto de dados reais. Apresentamos gráficos de resíduos padronizados [3] e alavanca generalizados [8].

2 Material e métodos

Modelos lineares simétricos tem assumido ligação identidade entre a média da variável resposta e um conjunto de preditores lineares. A classe de MLGS, introduz uma função de ligação entre a resposta média e o conjunto de preditores lineares, permitindo o uso de diferentes funções de ligação como por exemplo, ligação logarítmica, recíproca, dentre outras.

Assumimos que $Y_1,\ldots,Y_n$ são variáveis aleatórias independentes, em que cada Y_i tem uma distribuição exponencial potência com parâmetro de posição μ_i, parâmetro de escala ϕ e parâmetro de forma ν cuja função de densidade ([2]) é definida como

$$f(y_i,\mu_i,\phi)=\frac{1}{\sqrt{\phi}}\,C(\nu)\exp\left\{-\frac{1}{2}u_i^{1/(1+\nu)}\right\},\qquad -1<\nu\leqslant 1,\ u_i>0,\ , \tag{1}$$

em que $u_i=(y_i-\mu_i)^2/\phi$, $C(\nu)^{-1}=\Gamma(\frac{3+\nu}{2})2^{(3+\nu)/2}$ e cuja notação será $Y_i\sim \mathrm{EP}(\mu_i,\phi,\nu)$, $i=1,\ldots n$. Seja $\eta_i=x_i^\top\beta=\beta_1x_{i1}+\ldots+\beta_px_{ip}$ o preditor linear , em que $x_i=(x_{i1},\ldots,x_{ip})^\top$ é o

[1] LCE - ESALQ/USP. e-mail: *clobos@usp.br*
[2] Agradecimento à FAPESP pelo apoio financeiro.

vetor $p \times 1$ de covariáveis para o i-ésimo caso e $\beta = (\beta_1, \ldots, \beta_p)^\top$ é o vetor de parâmetros $p \times 1$. Assumimos que o parâmetro μ_i satisfaz

$$g(\mu_i) = \eta_i, \tag{2}$$

com $g(\cdot)$ sendo a função de ligação, monótona e pelo menos duas vezes diferenciável como em MLG. O modelo definido pelas equações (1)-(2) é chamado de MLGEP. O logaritmo da função de verossimilhança para o MLGEP pode ser escrito como

$$\ell(\theta) = -\frac{n}{2}\log(\phi) - n\log\left(\Gamma\left(\frac{3+\nu}{2}\right)\right) - \frac{n(3+\nu)}{2}\log(2) - \frac{1}{2}\sum_{i=1}^{n} u_i^{1/(1+\nu)},$$

em que $\theta = (\beta^\top, \phi)^\top$ (ν fixo). Usamos o método de máxima verossimilhança para estimar os parâmetros do modelo. Segue desde [8] que a matriz de alavanca generalizada é definida como,

$$GL(\theta) = D_\theta L_{\theta\theta}^{-1} L_{\theta y}, \tag{3}$$

avaliada em $\hat{\theta}$, com $D_\theta = \partial\mu/\partial\theta^\top = (D(a)X, 0)$, e $L_{\theta y} = \partial^2\ell(\theta)/\partial\theta\partial y^\top = (L_{\beta y}^\top, L_{\phi y}^\top)^\top$, em que $L_{\beta y} = X^\top D(c)D(a)/\phi$ e $L_{\phi y} = \partial^2\ell(\theta)/\partial\phi\partial y^\top = b^\top/\phi^2$. Então, depois de algumas manipulações algébricas, (3) pode ser escrito como

$$GL(\widehat{\theta}) = \widehat{H}_1 + \widehat{H}_2\widehat{H}_1 - \widehat{H}_2, \tag{4}$$

em que $H_1 = D(a)X\{X^\top D(d)X\}^{-1}X^\top D(a)D(c)$, $H_2 = \dfrac{1}{\phi^3 l_{\phi\phi\cdot 1}} D(a)X\{X^\top D(d)X\}^{-1}X^\top D(a)bb^\top$, e $l_{\phi\phi\cdot 1} = l_{\phi\phi} - L_{\phi\beta}L_{\beta\beta}^{-1}L_{\beta\phi}$. Note que para o modelo linear normal e ligação identidade $GL(\widehat{\theta}) = X(X^\top X)^{-1}X^\top$. Agora se ϕ for conhecido $L_{\beta\phi} \approx 0$, desde a equação (4) segue que

$$GL(\widehat{\beta}) = D(\widehat{a})X\{X^\top D(\widehat{d})X\}^{-1}X^\top D(\widehat{a})D(\widehat{c}) = \widehat{H}_1. \tag{5}$$

O gráfico índice de $\widehat{GL}_{ii}$ é utilizado para avaliar observações com uma alta influencia nos valores preditos. Por outro lado, para o modelo linear normal e com base em [3], podemos definir os resíduos padronizados para o MLGEP como

$$t_{r_i} = \frac{y_i - \hat{y}_i}{\sqrt{\hat{\phi}\xi_h\{1 - (4a_h\xi_h)^{-1}\hat{h}_{ii}\}}} \qquad i = 1, \ldots, n,$$

em que $a_h = \frac{\Gamma\{(3-\nu)/2\}}{4(2^{\nu-1})(1+\nu)^2\Gamma\{(\nu+1)/2\}}$, $\xi_h = -2\psi'(0)$, h_{ii} são os elementos da diagonal da matriz $H = W(a)^{1/2}X\{X^\top W(a)X\}^{-1}X^\top W(a)^{1/2}$, isto é, $h_{ii} = \omega(a_i)x_i^\top\{X^\top W(a)X\}^{-1}x_i$. Para n grande é fácil mostrar que $H_1 \cong H$. Para motivar nossa teoria usamos como exemplo o conjunto de dados de preços de casas que tem sido analisados por vários autores (veja, por exemplo, [1]).

Tabela 1: Resumo do ajuste do MLGS para os dados de preços de casas.

Distribuição	Preditor linear	Função de ligação	ν	AIC	BIC
Normal	quadrático	identidade	-	-38.8	-21.9
Exponencial Potência			0.4	-53.7	-36.8
Normal	quadrático	logarítmica	-	-40.2	-23.3
Exponencial Potência			0.4	-55.6	-38.7
Normal	quadrático	recíproca	-	-41.2	-24.3
Exponencial Potência			0.4	**-57.3**	**-40.4**

O objetivo do estudo é avaliar a associação entre os preços das casas com a qualidade do ar usando modelos de regressão. A variável resposta *LMV* (logaritmo do preço mediano das casas) está relacionado com 14 covariáveis. Para o nosso trabalho somente estudaremos a covariável *LSTAT* (logaritmo da proporção com renda mais baixa).

3 Resultados e discussões

A figura 1(esquerda) mostra o gráfico de dispersão com tendência não linear entre as variáveis *LMV* versus *LSTAT*.

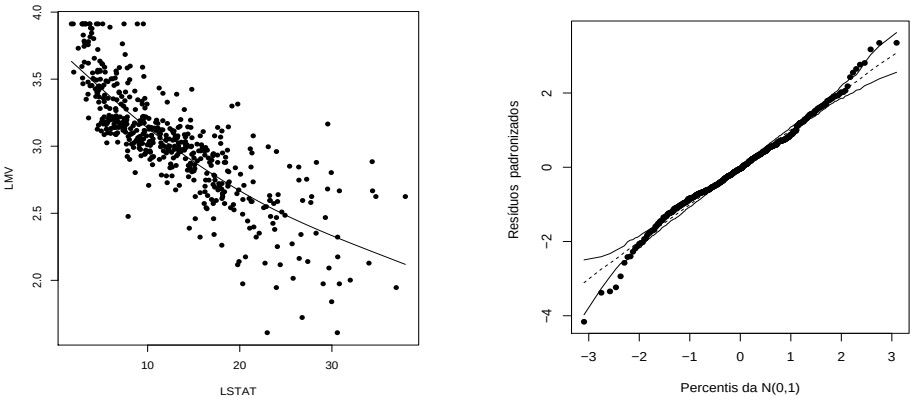

Figura 1: Gráfico de dispersão entre *LMV* e *LSTAT* (esquerda) e o QQplot com envelope simulado para o modelo linear quadrático com função de ligação recíproca(direita)

Assim, como sugerido pelo gráfico de dispersão (figura 1 (esquerda)), ajustamos os modelos: (i) $Y_i \overset{ind}{\sim} N(\mu_i, \phi)$ e $g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$ e as funções de ligação identidade, logarítmica e recíproca, em que $x = LSTAT$. A tabela 1 resume os valores para o *AIC* e *BIC* para cada ajuste, em que $AIC = -2\ell(\widehat{\theta}) + 2(p+1)$ e $BIC = -2\ell(\widehat{\theta}) + (p+1)\log(n)$. Notamos que o melhor ajuste (sob erros normais) é atingido para o preditor quadrático com ligação recíproca. No entanto, o gráfico QQplot para os resíduos t_{r_i} (figura 1 (direita)) indica que esse modelo não ajusta bem os dados, sugerindo o uso de uma distribuição simétrica com caudas mais pesadas do que a distribuição normal. Assim, consideramos como alternativa a distribuição exponen-

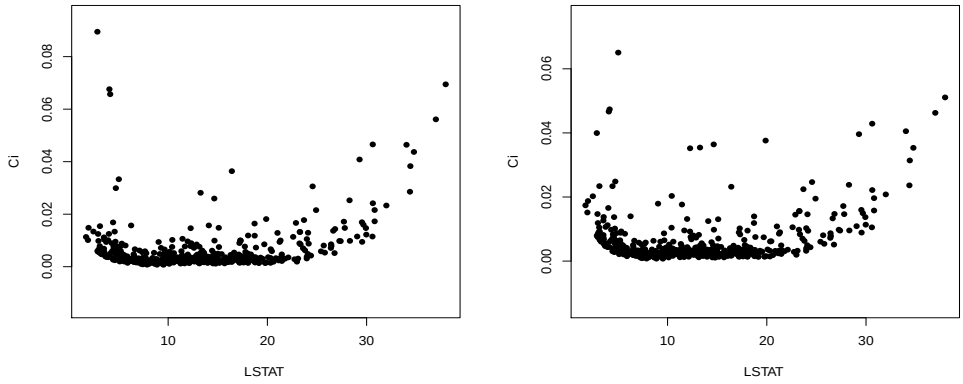

Figura 2: Gráfico entre C_i e *LSTAT* para o modelo linear generalizado exponencial potência com preditor linear quadrático e ligações logarítmica (esquerda) e recíproca (recíproca) sob ponderação de casos.

cial potência, isto é, $Y_i \overset{ind}{\sim} \text{EP}(\mu_i, \phi, \nu)$ com $g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$. Os resultados são resumidos na tabela 1. O parâmetro ν foi selecionado para cada modelo usando o critério *AIC*. Notamos que o melhor ajuste foi obtido para o modelo linear generalizado exponecial potência com preditor linear quadrático e função de ligação recíproca. A figura 2 mostra o gráfico entre C_i, para o caso de ponderação de casos, e *LSTAT* para o modelo linear generalizado exponecial potência com preditor linear quadrático e ligação recíproca. Observamos uma alta sensibilidade de C_i para valores altos de *LSTAT*, o que significa que a predição do preço mediano das casas parece ser mais difícil para a proporção da renda mais baixa. O gráfico entre $\hat{h}_{ii}$ e *LSTAT* (omitido aqui) apresenta a mesma tendência e o QQplot para os resíduos t_{r_i} com envelope simulado (figura 3) não indica presença de observações atípicas nem afastamento das suposições do modelo.

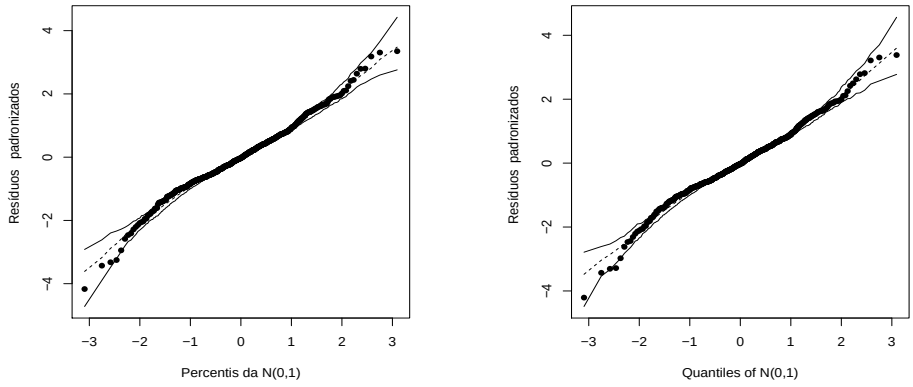

Figura 3: QQplot com envelopes simulados para t_{r_i} para o modelo linear generalizado exponencial potência com preditor linear quadrático e ligações logarítmica (esquerda) e recíproca (recíproca) para os preços das casas.

As estimativas de máxima verossimilhança (erro padrão) do modelo selecionado são dadas por $\hat{\beta}_0 = 0.257057(0.00330906)$, $\hat{\beta}_1 = 0.00704961(0.000555751)$ $\hat{\beta}_2 = -0.0000581289(0.0000193211)$ e $\hat{\phi} = 0.0248689(0.00184995)$. Então, o valor predito tem a seguinte forma

$$\hat{\mu}(x) = (0.257057 + 0.00704961x - 0.0000581289x^2)^{-1},$$

em que $\mu(x)$ denota os valores esperados de LMV dado $x = LSTAT$.

4 Conclusões

Em este trabalho apresentamos o modelo linear generalizado exponencial potência como alternativa ao modelo normal para diferentes funções de ligação seguindo a ideia dos modelos lineares generalizados. Além do anterior, apresentamos alguns gráficos de resíduos e alavanca generalizados para o modelo proposto com base num conjunto de dados reais.

Referências

[1] BELSEY, D. A.; KUH, E.; WELSCH, R. E. (1980). *Regression Diagnostics. Identifying Influential Data and Sources of Collinearity.* Wiley, New York.

[2] BOX, G. E. P.; TIAO, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Reading, Addison-Wesley.

[3] COX, D. R.; SNELL, E. J. (1968). A general definition of residuals, *Journal of the Royal Statistical Society, B, 30, 2, 248-275*.

[4] DEVROYE, L. (1986). *Non-Uniform Random Variable Generation.* Springer-Verlag, New York.

[5] GALEA, M.; PAULA, G. A.; URIBE-OPAZO M. (2003). On influence diagnostics in univariate elliptical linear regression models. *Statistical Papers, 44, 23-45*.

[6] MCCULLAGH, P.; NELDER, J. A. (1989). *Generalized Linear Models*, 2nd Edition. Chapman and Hall, London.

[7] ST. LAURENT, R. T.; COOK, R. D. (1992). Leverage and superleverage in nonlinear regression, *Journal of the American Statistical Association, 87, 985-990*.

[8] WEI, B. C.; HU, Y. Q.; FUNG, W. K. (1998). Generalized leverage and its applications, *Scandinavian Journal of Statistics, 25, 25-37*.