ASSOCIAÇÃO ENTRE OS DELITOS REGISTRADOS NA RGV: UMA APLICAÇÃO DO MODELO GLARMA POISSON

Alyne Neves Silva[1], Valdério Anselmo Reisen[2]

Resumo: *Neste trabalho é utilizado o modelo Poisson GLARMA para descrever os delitos comumentes registrados na Região da Grande Vitória, ES, Brasil, pelo CIODES. Com base no trabalho de Davis et al. (1999), que provém uma revisão de modelos para séries temporais com distribuição de Poisson, procedimentos de identificação e análise residual do modelo GLARMA são considerados. Os resultados do ajuste indicaram que o modelo GLARMA é bastante apropriado para modelar os dados observados.*

Palavras-chave: Modelo Linear Generalizado, GLARMA, Modelo ARMA.

Introdução

Nos últimos anos ocorreu um considerável aumento no estudo de modelos de séries temporais para variáveis com distribuição de probabilidade não-Gaussiana. Grande parte desses estudos considera a variável de interesse como sendo um processo discreto. Mais precisamente, a série temporal de interesse é a contagem de determinado evento, que ocorre num dado período de tempo em uma taxa média conhecida, e cada evento observado é independente do tempo decorrido. Nesse caso, a distribuição de probabilidade "candidata natural" é a Poisson.

O Modelo Linear Generalizado (MLG), proposto por [12], foi uma das primeiras metodologias apresentadas sobre o estudo de modelagem para dados discretos. Este pode ser interpretado como uma generalização do tradicional modelo de regressão linear. Anos à frente, [11] formalizaram as idéias que envolviam a estrutura teórica, os procedimentos de estimação e os métodos de adequação do MLG.

Inúmeros trabalhos relacionados ao MLG foram realizados desde 1972, pode-se citar [15], e [7], entre outros.

Contudo, essa metodologias não considera, como em um modelo de regressão, as relações que podem ocorrer entre as observações analizadas ao longo do tempo. Dessa forma, surge a necessidade e o interesse em se realizar estudos que combinassem os métodos de regressão e de série temporais. [3] define em seu trabalho duas classes de modelos para análise de séries temporais não-Gaussianas, os *observation-driven models* e os *parameter-driven models*. No *parameter-driven model* existe um processo latente que rege a função média condicional. Já no *observation-driven model* a estrutura de dependência é introduzida através da incorporação dos valores desfasados das contagens observadas, obtidas diretamente da função média do modelo.

No contexto do trabalho de [3, 16, 17, 10, 9, 14, 6, 4, 5, 1], entre outros autores abordaram o estudo de modelos matemáticos para análise de séries temporais discretas. Os autores acima citados, derivam em seus trabalhos a metodologia clássica de séries temporais ARMA(p,q) [2],

[1]Estatística Aplicada e Biometria, PPESTBIO, UFV,
alyne.silva@ufv.br
[2]Departamento de Estatística, UFES,
valderioanselmoreisen@gmail.com

de forma a unificar aos modelos de regressão generalizada as componentes autorregressivas e médias móveis.

Na classe do *observation-driven model* destaca-se a interessante metodologia do modelo GLARMA (*Generalized Linear Autoregressive Moving Average Models*), proposto primeiramente por [14] e, sequencialmente apresentado por [6, 4, 5]. O modelo GLARMA estende a estrutura familiar do MLG, de forma a permitir a correlação serial entre as observações, bem como uma variação binomial extra nos dados, e obter o logaritmo natural da média condicional do processo como uma função linear das observações passadas.

O modelo GLARMA é utilizado para modelar uma variedade de variáveis respostas dependentes do tempo (que também sejam covariáveis dependentes do tempo), que possuam distribuição marginal condicional pentencente à família exponencial, por exemplo, dados contínuos com distribuição Gama condicional (e.g., a volatilidade no modelo GARCH) ou dados de contagem com distribuição condicional binomial negativa, binomial ou Poisson.

Neste trabalho a metodologia do GLARMA Poisson é utilizada para modelar o *número diário de delitos registrados na Região da Grande Vitória (RGV)*. Os dados de delitos são referentes aos principais crimes contra a pessoa e contra o patrimônio registrados nos municípios da Grande Vitória pela Gerência de Estatística e Análise Criminal da Secretaria de Estado da Segurança Pública e Defesa Social do Espírito Santo.

Modelo Linear Generalizado

Trata-se da generalização do modelo de regressão linear e gaussiano de forma a adequá-lo para a modelagem de variáveis de resposta independentes que apresentem características explícitas de não-normalidade, tais como variáveis contínuas com assimetria e dados de contagem.

O Modelo Linear Generalizado, segundo [11], pode ser caracterizado em três pontos:

1. **Componente Aleatória.** Considerem-se N variáveis aleatórias Y_i (i = 1,...., n) independentes, de média μ_i respectivamente e função de probabilidade ou função densidade de probabilidade pertencente à família exponencial, isto é

$$f(y_i|\theta_t, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\alpha_i(\phi)} + c(y_i, \phi)\right\} \tag{1}$$

   em que $a_i(.)$, $b(.)$ e $c(.)$ são funções específicas para cada distribuição. Se ϕ for conhecido tem-se uma distribuição da família exponencial com parâmetro canônico θ.

2. **Componente Sistemática.** As N observações destas p variáveis constituem a matriz $\mathbf{X}$ (i = 1, ..., n).

   A partir desta matriz define-se um preditor linear η_i , i = 1,...,n, da forma

$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$$

   constituindo os β_j, j = 1,...,p, um vetor de parâmetros desconhecidos, a estimar a partir dos dados.

3. As duas componentes anteriores relacionam-se através de uma função de ligação g_i, que se admite existir, ser monótona e diferenciável, e que transforma μ_i em η_i, ou seja

$$\eta_i = g_i(\mu_i), i = 1, ..., N. \tag{2}$$

Essa função de ligação é invertível, logo é possível obter

$$g^{-1}(\eta_i) = \mu \tag{3}$$

que é denominada função média.

Ficam assim definidas as componentes do MLG. Suponha agora que a variável de interesse, Y_i, segue um processo de Poisson com média μ_i com função densidade

$$P(Y_i = y_i) = \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!},\ y_i = 0, 1, 2, \ldots. \tag{4}$$

Reescrevendo (4) na forma de (1)

$$\begin{aligned} P(Y_i = y_i) &= \exp\left\{\log\left[\frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}\right]\right\} \\ &= \exp\left\{\log\left[\exp(-\mu_i)\mu_i^{y_i}\right] - \log(y_i!)\right\} \\ &= \exp\{-\mu_i + y_i\log(\mu_i) - \log(y_i!)\} \\ &= \exp\{y_i\log(\mu_i) - \mu_i - \log(y_i!)\}. \end{aligned} \tag{5}$$

Comparando o resultado (5) com (1), pode-se concluir que $a_i(\phi) = 1$, $b(\theta_i) = \log(\mu_i)$ e $c_i(y, \phi) = -\log(y!)$. Fazendo $\theta_i = \log(\mu_i)$, então tem-se que $\mu_i = \exp(\theta_i)$.

Assim, considerando o vetor de covariáveis (regressoras) $\mathbf{x}$

$$\begin{aligned} \mu_i = \exp(\theta_i) &\implies g^{-1}(\eta_i) = \mu_i = \exp(\theta_i) \\ &\implies \log(\mu_i) = \theta_i \implies \eta_i = \log(\mu_i) \\ &\implies \log(\mu_i) = \mathbf{x}_i\beta,\ i = 1, \ldots, n. \end{aligned} \tag{6}$$

Quando a variável de interesse é um processo de contagem, isto é, um processo de Poisson, o modelo linear generalizado obtido é referido apenas como *modelo de regressão de Poisson*, por ser derivado da parametrização da relação entre o parâmetro μ, média, e as covariáveis ou regressoras. De acordo com (6), a suposição padrão é utilizar a parametrização da média exponencial,

$$\mu_i = \exp(\mathbf{x'}_i\beta),\ i = 1, \ldots, n. \tag{7}$$

A estimação do MLG dá-se pelo método de máxima verossimilhança.

Modelo GLARMA

Esta seção está fundamentada em descrever o modelo Autorregressivo Média Móvel Linear Generalizado (GLARMA) de acordo com o trabalho apresentado por [6, 4, 5].

O GLARMA é uma combinação dos modelos MLG e ARMA, [2] , sendo considerado uma extensão para distribuições condicionais não-Gaussianas, pertencentes à família exponencial. Este é defenido neste trabalho com a mesma notação utilizada nos MLG para amostras independentes.

Sejam $\{Y_t\}$ e $\mathbf{Z}_{t-1} = (Z_{(t-1)1}, ..., Z_{(t-1)p})$, para cada $t = 1, ..., n$, a série temporal de interesse e o correspondente vetor p-dimensional do passado das variáveis explanatórias ou covariáveis, respectivamente. Seja $\Im$ o campo-σ gerado por $Y_{t-1}, Y_{t-2}, ..., \mathbf{Z}_{t-1}, \mathbf{Z}_{t-2}, ...$, isto é, valores passados da série resposta e possíveis valores do presente (quando conhecidos) das covariáveis

$$\Im_{t-1} = \sigma\{Y_{t-1}, Y_{t-2}, ..., \mathbf{Z}_{t-1}, \mathbf{Z}_{t-2}, ...\}.$$

Séries temporais segundo o MLG podem ser definidas com as seguintes modificações nas componentes aleatória e sistemática, [8].

1. A componente aleatória, apresentada em (1), passa ser definida a partir da distribuição condicional da resposta dado o passado, isto é, para $t = 1,...,$ n

$$f(y_t; \theta_t, \phi|\Im_{t-1}) = \exp\left\{\frac{y_t\theta_t - b(\theta_t)}{\alpha_t(\phi)} + c(y_t, \phi)\right\}; \tag{8}$$

2. A componente sistemática passa a ser da seguinte forma

$$g(\mu_t) = \eta_t = \mathbf{x}_t'\beta + Z_t = \mathbf{x}_t'\beta + \sum_{i=1}^{\infty}\gamma_i e_{t-i} \tag{9}$$

onde $e_t = (Y_t - \mu_t)/\mu_t^{\lambda}$, $\lambda \in (0, 1]$, é uma sequência diferença martingale e γ é o vetor de parâmetros.

Conforme apresentado na seção anterior, quando a variável de interesse, Y_t, segue um processo de Poisson com média μ_t, tem-se que $\eta_t = g(\mu_t) = \log(\mu)$. Assim, no contexto deste trabalho de modelos de regressão para séries temporais de contagem, assume-se que

$$Y_t|\Im_{t-1} \sim Po(\mu_t).$$

Logo, o processo $\log(\mu_t)$ é dirigido por um ruído que é uma sequência diferença martingale gerada pelo conjunto de dados observado.

De acordo com [5] é possível especificar o termo média móvel infinito neste modelo por um número finito de parâmetros. Mais precisamente,

$$\sum_{i=1}^{\infty}\gamma_i e_{t-i} = \sum_{i=1}^{\infty}\gamma_i z^i = \theta(z)/\phi(z) - 1,$$

onde $\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p$ e $\theta(z) = 1 + \theta_1 z + \ldots + \theta_q z^q$ são, respectivamente, os polinômios autorregressivos e média móvel do filtro ARMA, cada um possuindo seus zeros fora do círculo unitário, e γ é o vetor dos parâmetros consistindo nesses $\phi_i's$ e $\theta_j's$. Assim, segue que $\{Z_t\}$ pode ser computado como nas recursões do modelo autorregressivo média móvel

$$Z_t = \sum_{i=1}^{p}\phi_i(Z_{t-i} + e_{t-i}) + \sum_{i=1}^{q}\theta_i e_{t-i}. \tag{10}$$

Propriedades do modelo

Seja

$$W_t = \log(\mu_t) = \mathbf{x}_t'\beta + Z_t,$$

segue que, inicialmente, $e_s = 0$ e $Y_s = 0$ para $s \leq 0$, $\Im_{s-1}^e = \{e_t : t \leq s-1\}$ e $\Im_{s-1} = \{Y_t : t \leq s-1\}$ geram o mesmo campo-σ e, como definido anteriormente,e_t forma uma sequência diferença martingale,

$$E(e_s|\Im_{s-1}^e) = 0, \ para \ s \geq 1,$$

onde $\Im_e^{s-1}$ é o σ-álgebra gerado por $\{e_t : t \leq s-1\}$. Como e_t tem média zero, sua variância é

$$Var(e_t) = E(e_t^2) = E[E(e_t^2|\Im_{t-1})] = \mu_t^{1-2\lambda},$$

que é unitário quando $\lambda = 0,5$. Outra propriedade da diferença martingale é que a covariância, para $s \neq t$, é

$$E(e_t e_s) = 0.$$

Dessas propriedades segue que, para qualquer λ,

$$E(W_t) = \mathbf{x}_t'\beta \qquad \text{e}$$

$$Var(W_t) = \sum_{i=1}^{\infty} \gamma_i^2 \mu_{t-i}^{1-2\lambda}, \text{ e, para } l > 0,$$

$$Cov(W_t, W_{t+l}) = \sum_{i=1}^{\infty} \gamma_i \gamma_{i+l} \mu_{t-i}^{1-2\lambda},$$

e, novamente, se $\lambda = 0,5$, as covariâncias não dependem do tempo t, nem mesmo se $\{\mu_t\}$ não for estritamente estacionária.

Estimação

[6, 4, 5] estabelecem uma aproximação da função de verossimilhança similar as apro-ximações usadas nos modelos de séries temporais lineares. No entanto, as propriedades de estimação e inferência para o modelo são consideradas somente quando $\lambda = 0,5$.

Para estimar os parâmetros do modelo GLARMA a função de verossimilhança deve ser maximizada. A aproximação da verosssimilhança, com as derivadas de primeira e segunda ordem, podem ser calculadas recursivamente utilizando o procedimento de Newton-Raphson. De acordo com [12] as estimativas dos parâmetros do MLG pelo método de Mínimos Quadrados Iterativos Reponderados (*Iteratively Reweighted Least Squares* - IWLS) são equivalentes as estimativas de máxima verossimilhança, levando em consideração algumas suposições. A função `glm` do **R** [13] utiliza o método IRLS para obter os parâmetros estimados.

Outra ferramenta utilizada na estimação dos parâmetros autorregressivos e média móvel, p e q respectivamente, do GLARMA é a identificação do modelo à partir das funções de autocorrelação. A Função de Autocorrelação (FAC) e a Função de Autocorrelação Parcial (FACP) para $Y_t - g^{-1}(\mathbf{x}_t\beta)$ são apropriadas para identificação do modelo.

Diagnósticos

O diagnóstico de um modelo de regressão consiste em explorar e testar a adequacidade e a qualidade do ajuste do modelo estimado. Nos modelos lineares generalizados, isso cocorre a partir de uma análise dos resíduos e do *deviance*.

Para séries temporais de contagem $\{Y_t\}$, sob um modelo *log*-linear de Poisson, o *deviance* tem a forma

$$Deviance = -2 \sum_{t=1}^{n} \left\{ Y_t \log\left(\frac{Y_t}{\hat{\mu}_t}\right) - (Y_t - \hat{\mu}_t) \right\}.$$

O *deviance* é comumente associado aos critérios de informação de Akaike (AIC) e Bayesiano (BIC). Estes são utilizados para avaliação e seleção de modelos.

Aplicação

O conjunto de dados observados trata-se do número diário de delitos ou ocorrências registradas nos municípios de Vitória, Vila Velha, Guarapari, Viana, Serra e Cariacica, no período de 01 de janeiro de 2005 a 25 de maio de 2007. Os dados foram contabilizados pelo Centro Integrado Operacional de Defesa Social (CIODES), compreendendo um total de 875 dias. Realizou-se a estimação das componentes de séries temporais do modelo GLARMA(p,q) Poisson para cada um dos municípios em que foram registrados a presença de autocorrelação. As análises foram realizadas na linguagem computacional **R**, [13], considerando nível de significância de 5%.

A Tabela 1 apresenta o AIC, o BIC e a variância estimada, $\hat{\sigma}^2$, dos modelos estimados. Com base nesses resultados é evidente que para os municípios de Cariacica, Guarapari, Viana e Vitória a estrutura do GLARMA Poisson identificada foi o GLARMA(1,1). Para o município de Vila Velha, os valores identificados pelo AIC e BIC para os modelos estimados divergiram. Logo,

considerando o princípio da parcimônia, optou-se na escolha do modelo que possuisse a menor quantidade de parâmetros, isto é, os delitos registrados no município de Vila Velha podem ser estimados pelo modelo GLARMA(1,1).

Conclusões

O objetivo principal deste trabalho foi avançar no estudo da modelagem dos processos de séries espaço-temporal e regressão generalizada considerando casos onde a variável de interesse ou resposta seja não Gaussiana (normal), mais precisamente, possua distribuição de Poisson. Ela expressa, por exemplo, a probabilidade de certo número de eventos ocorrerem num dado período tempo, caso estes ocorram com uma taxa média conhecida e caso cada evento seja independente do tempo decorrido desde o último evento.

Como metodologia de modelagem utilizou-se o conceito do Modelo Autorregressivo Média Móvel Linear Generalizado, o GLARMA. A forma funcional do GLARMA, propriedades, estimação e diagnósticos foi apresentada por [4] e esta que se utiliza neste trabalho

Realizando o estudo de regressão para análise de séries constatou-se que o modelo Poisson GLARMA adequou-se muito bem para descrever o número diário de delitos registrados nos município da RGV. Com exceção dos municípios de Viana e Serra que, devido a existência de vários dias em que não foram registrados delitos nesses municípios (presença de valores nulos), os pressupostos para ajuste dos modelos não foram satisfatórios. Pelos ajustes, verificou-se que os delitos dos municípios podem ser descritos pelo modelo Poisson GLARMA (1,1). De acordo com esse modelo, os delitos que são registrados em um dia qualquer, tem relação com os delitos que foram registrados um dia atrás. Além disso, o mesmo delito só é registrado um dia após o mesmo ter ocorrido.

Referências

[1] BENJAMIN, R. A.; RIGBY M. A.; STASINOPOULOS, M. D., Generalized autoregressive moving average models. *Journal of the American Statistical Association*, **98(461)**, 214-223, 2003.

[2] BOX, G. E. P.; JENKINS, G. M., *Time series analysis: Forecasting and Control*, Holden Day, 1976.

[3] COX, D. R., Statistical analysis of time series: Some recent developments. *Scandinavian Journal of Statistics*, **8**, 93-115, 1981.

[4] DAVIS, R. A.; DUNSMUIR,W. T. M.; WANG, Y., Modelling time series of count data. In S. Ghosh, editor, *Asymptotics, Nonparametric & Time Series*, pp. 63- 114. Marcel Dekker, New York, 1999.

[5] DAVIS, R. A.; DUNSMUIR,W. T. M.; WANG, Y., On autocorrelation in a Poisson regression model. *Biometrika*, **87**, 491-505, 2000.

[6] DAVIS, R. A.; DUNSMUIR, W. T. M.; STREETT, S. B., Observation-driven models for Poisson counts, *Biometrika*, 90(4), 777-790, 2003.

[7] JØRGENSEN, B., Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models. *Biometrika*, **70**, 19-28, 1983.

[8] KEDEM, B.; FOKIANOS, K., *Regression Models for Time Series Analysis*. Wiley, USA, 2nd, 2002.

[9] LI, W. K., Time series models based on generalized linear models: Some further results. *Biometrics*, **50**, 506-511, 1994.

[10] McKENZIE, E., Some ARMA models for dependent sequences of Poisson counts. *Advances in Applied Probability*, **(20)**, 822 - 835, 1988.

[11] McCULLAGH, P.; NELDER, J. A., *Generalized Linear Models*. Chapman and Hall, Londres, 2nd, 1989.

[12] NELDER, J. A.; WEDDERBURN, R. W., Generalized linear models. *Journal of the Royal Statistical Society Series A*, **135(3)**, 370-384, 1972.

[13] R DEVELOPMENT CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.r-project.org, 2010.

[14] SHEPHARD, N., *Generalized Linear Autoregressions*. (Unpublished paper, Oxford University, UK), 1995.

[15] WEDDERBURN, R. W. M., Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika*, **61**, 439-447, 1974.

[16] ZEGER, S.L., A regression model for time series of counts. *Biometrika*, **75 4**, 621-629, 1988.

[17] ZEGER, S. L.; LIANG, K. Y., Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**, 121-130, 1986.

Tabela 1: Critérios AIC, BIC e variância estimada para o Modelo GLARMA Poisson

Município		AIC				BIC				$\hat{\sigma}^2$			
		p											
	q	0	1	2	3	0	1	2	3	0	1	2	3
Cariacica	0		2.562,49	2.558,81	2.559,52		2.563,25	2.564,35	2.569,84		1,09	1,08	1,08
	1	2.563,25	**2.555,92**	2.557,88		2.564,03	**2.561,47**	2.568,20		1,09	1,08	1,08	
	2	2.559,61	2.557,88	2.559,18		2.565,17	2.568,20	2.574,28		1,08	1,08	1,08	
	3	2.560,19	2.558,80	2.561,18		2.570,52	2.573,90	2.581,05		1,08	1,08	1,08	
Guarapari	0		2.785,77	2.762,71	2.751,02		2.786,55	2.768,27	2.761,34		1,40	1,36	1,34
	1	2.796,72	**2.721,79**	2.723,58		2.797,49	**2.727,35**	2.733,90		1,42	1,30	1,30	
	2	2.779,68	2.723,57	2.725,48		2.785,23	2.733,90	2.740,58		1,39	1,30	1,30	
	3	2.770,94	2.725,30	2.727,46		2.781,26	2.740,40	2.747,33		1,37	1,30	1,30	
Viana	0		2.807,52	2.778,70	2.762,39		2.808,29	2.784,25	2.772,70		1,44	1,39	1,36
	1	2.824,48	**2.728,81**	2.730,75		2.825,25	**2.734,37**	2.741,06		1,47	1,31	1,31	
	2	2.803,21	2.730,74	2.732,70		2.808,75	2.741,06	2.747,80		1,43	1,31	1,31	
	3	2.790,56	2.732,36	2.734,70		2.800,88	2.747,46	2.754,57		1,41	1,31	1,31	
Vila Velha	0		2.664,71	2.659,90	2.656,34		2.665,47	2.665,45	2.666,66		1,22	1,21	1,21
	1	2.669,08	2.627,64	2.627,00		2.669,85	**2.633,19**	2.637,32		1,23	1,17	1,17	
	2	2.664,47	**2.626,96**	2.628,96		2.670,01	2.637,28	2.644,06		1,22	1,17	1,17	
	3	2.662,11	2.628,93	2.630,87		2.672,42	2.644,04	2.650,75		1,21	1,17	1,17	
Vitória	0		2.611,88	2.607,45	2.609,44		2.612,65	2.612,99	2.619,76		1,15	1,14	1,14
	1	2.612,08	**2.595,32**	2.609,22		2.612,85	**2.600,87**	2.619,54		1,15	1,13	1,14	
	2	2.607,95	2.609,94	2.596,27		2.613,49	2.620,26	2.611,36		1,14	1,14	1,12	
	3	2.609,94	2.598,20	2.597,84		2.620,26	2.613,30	2.617,71		1,14	1,13	1,12	