Modelos INAR: Estimação de parâmetros e suas Propriedades em amostras finitas

Palavras-Chave: Dados de contagem, Estimação, Previsão

Autores(as):

Nathaly Lissa Shinosaki Izawa, IMECC – UNICAMP

Prof. Dr. Carlos Trucíos, IMECC – UNICAMP

INTRODUÇÃO

Nas últimas décadas, séries temporais que envolvem dados de contagem vêm ganhando cada vez mais importância e popularidade. Isto é devido à presença deste tipo de dados em diversas áreas de conhecimento. Por exemplo, alguns casos práticos de séries temporais que envolvem este tipo de dados podem ser encontrados em economia, finanças, previsão da demanda, meio ambiente, experimentos biológicos, saúde pública, seguros, medicina, entre outros **(Turkman, Scotto, & Bermudez, 2014; Pavlopoulos & Karlis, 2008; Silva, 2015)**.

A maioria dos modelos estudados em cursos introdutórios de séries temporais não são específicos para dados de contagem. Contudo, muitas das séries temporais nas quais estamos interessados no dia a dia são séries temporais que possuem esta característica. Assim, como uma alternativa para lidar com séries temporais para dados de contagem, surge o modelo INAR (*Integer-valued Autoregressive*) descrito pôr Alzaid e Al-Osh **(Alzaid & Al-Osh, 1987)**, sendo este modelo análogo ao modelo ARMA (*Autoregressive-moving-average*).

Deste modo, este projeto teve como intuito explorar do modelo INAR, entender como ele funciona tanto na parte teórica, testando alguns métodos de estimação para seus parâmetros, quanto na parte aplicada, fazendo simulações através de programas computacionais apropriados e implementando o método de previsão. Por motivos de simplicidade e de tempo, focamos no modelo INAR(1) com o erro com uma distribuição Poisson(λ), contudo é também possível trabalhar com ordens $\boldsymbol{p}$ mais gerais, bem como outras distribuições para os erros.

METODOLOGIA

Para este projeto, foi feito uma revisão bibliográfica de diversos artigos e livros, listados na Bibliografia, que exploraram o modelo INAR para uma melhor compreensão dele, como é feito a estimação de seus parâmetros α e λ, e uma maneira de fazer previsões para a série temporal analisada.

Considere Y um valor aleatório não-negativo e $\alpha \in [0,1]$, o operador "$\circ$" **(Harn & Steutel, 1979)** é dado por:

$$\alpha \circ Y = \sum_{i=1}^{Y} x_i$$

Onde $\{x_i\}$ é uma sequência de variáveis aleatórias independentes e identicamente distribuídas (iid), independentes de Y, sendo

$$Pr(x_i = 1) = 1 - Pr(x_i = 0) = \alpha$$

Então temos o *binomial thinning*, uma distribuição binomial com probabilidade α e tamanho Y. Dessa forma, o modelo INAR(1) é dado por:

$$Y_t = \alpha \circ Y_{t-1} + \epsilon_t,\ t \in \mathbb{Z}$$

Sendo ϵ_t o erro com qualquer distribuição com valores inteiros não-negativos e não-correlacionados, com média μ e variância σ^2.

Deste modo, utilizando a linguagem de programação R, disponível gratuitamente em https://www.r-project.org, foram escritos códigos para fazer a simulação de um INAR(1) com erro com distribuição Poisson(λ), com $\lambda = 1$.

Duas simulações de Monte Carlo com 10000 replicações foram feitas, uma sem *outliers* e outra com *outliers*, para cada um dos 3 métodos de estimação considerados, sendo eles:

1. Yule-Walker (YW)

   Considere que temos uma amostra $Y_1, \dots, Y_t$ de um processo $\{Y_t\}_{t\in\mathbb{Z}}$ estacionário, a Função de Autocorrelação (ACF) é dada por:

   $$\hat{\rho}(k) = \frac{\hat{\gamma}(k)}{\hat{\gamma}(0)} = \frac{\sum_{t=1}^{T-k}(Y_t - \bar{Y})\,(Y_{t+k} - \bar{Y})}{\sum_{t=1}^{T}(Y_t - \bar{Y})^2}$$

   Onde $\bar{Y} = 1/T \sum_{t=1}^{T} Y_t$ é a média da amostra. Como os estimadores de YW se baseiam no ACF amostral $\hat{\rho}(k)$, usando que $\rho_Y(1) = \alpha$ e o primeiro momento de Y_t, $\mathbb{E}(Y_t) = \lambda/(1-\alpha)$, os estimadores são dados por:

   $$\hat{\alpha}_{YW} = \hat{\rho}_{YW}(1) = \frac{\sum_{t=1}^{T-1}(Y_t - \bar{Y})\,(Y_{t+1} - \bar{Y})}{\sum_{t=1}^{T}(Y_t - \bar{Y})^2} \text{ e } \hat{\lambda}_{YW} = (1 - \hat{\alpha}_{YW})\bar{Y}$$

2. Squared Difference (SD)

   Os estimadores de SD **(Bourguignon & Vasconcellos, 2015)** são dados por:

   $$\hat{\lambda}_{SD} = \frac{1}{2(T-1)}\sum_{t=2}^{T}(Y_t - \bar{Y})^2 \text{ e } \hat{\alpha}_{SD} = 1 - \frac{\hat{\lambda}_{SD}}{\bar{Y}}$$

3. Kendall (K)

   O coeficiente de correlação de Kendall é usado para testes de independência entre duas variáveis, ou seja, sob hipótese nula, assumimos que os pares $(X_1, Y_1), \dots, (X_T, Y_T)$ de uma amostra aleatória bivariada são iid. Assim, estimador do coeficiente de correlação $\hat{\rho}_K(k)$ **(Kendall, 1938)**, baseado na estatística K, pode ser utilizado para estimar o $\rho(k)$ e é definido como

   $$\hat{\rho}_K = \frac{2K}{T(T-1)} \text{ onde } K = \sum_{i=1}^{T-1}\sum_{j=i+1}^{T} Q\left[(Y_i, Y_{i+k})(Y_j, Y_{j+k})\right]$$

   Com

$$Q[(a,b)(c,d)] = \begin{cases} 1, & \text{se } (d-b)(c-a) > 0 \\ 0, & \text{se } (d-b)(c-a) = 0 \\ -1, & \text{se } (d-b)(c-a) < 0 \end{cases}$$

Desta maneira, com $k = 1$, temos que

$$\hat{\alpha}_K = \hat{\rho}_K(1) = \frac{2K}{T(T-1)} \text{ e } \hat{\lambda}_K = (1-\hat{\alpha}_K)\bar{Y}$$

Nas mesmas simulações de Monte Carlo, também foram feitas a previsão dos dados simulados usando o alpha e o lambda simulados, e fazendo a probabilidade acumulada da probabilidade condicional usando os parâmetros estimados de cada método, ou seja,

$$P(Y_t = y | Y_{t-1}) \leq 0.5$$

Sendo assim, o valor da probabilidade condicional para $y = 0,1,2,\ldots$ até que a probabilidade acumulada seja menor ou igual a 0.5.

RESULTADOS E DISCUSSÃO

Os resultados das simulações de Monte Carlo, com e sem *outliers*, estão representados na Tabela 1 e na Tabela 2, respectivamente. Com as tabelas, nota-se que, em ambos os casos, os métodos de Yule-Walker e Squared Difference têm melhores resultados ao estimarem os parâmetros α e λ. Conforme o valor de α vai aumentando, o erro quadrático médio vai aumentando. No caso da simulação com *outliers*, percebemos que as estimações são piores do que as que não possuíam *outliers*, o que era esperado.

Tabela 1- Médias empíricas e o Erro quadrático médio (entre parênteses) das estimativas para os parâmetros α e λ, para valores de T, utilizando os métodos de Yule-Walker, Squared Difference e Kendall. Dados simulados com $\lambda = 1$ e os valores correspondentes de α,e T.

		$\alpha = 0.2$		$\alpha = 0.5$		$\alpha = 0.8$	
Método	T	α	λ	α	λ	α	λ
Yule-Walker	100	0.184	1.02	0.466	1.06	0.756	1.21
		(0.010)	(0.024)	(0.010)	(0.045)	(0.007)	(0.171)
	300	0.193	1.01	0.489	1.02	0.786	1.07
		(0.004)	(0.008)	(0.003)	(0.014)	(0.002)	(0.040)
	500	0.196	1.01	0.495	1.01	0.791	1.04
		(0.002)	(0.005)	(0.001)	(0.008)	(0.001)	(0.023)
Squared Difference	100	0.207	0.995	0.496	1000	0.799	1000
		(0.015)	(0.037)	(0.007)	(0.034)	(0.001)	(0.031)
	300	0.199	1000	0.500	0.998	0.799	0.999
		(0.006)	(0.012)	(0.002)	(0.011)	(0.000)	(0.010)
	500	0.200	1000	0.499	1000	0.800	1000
		(0.004)	(0.008)	(0.001)	(0.007)	(0.000)	(0.006)
Kendall	100	0.107	1.11	0.294	1.41	0.532	2.33
		(0.012)	(0.030)	(0.046)	(0.202)	(0.076)	(1.95)
	300	0.112	1.11	0.309	1.38	0.563	2.18
		(0.009)	(0.018)	(0.038)	(0.156)	(0.058)	(1.45)
	500	0.113	1.11	0.313	1.37	0.569	2.15
		(0.008)	(0.015)	(0.036)	(0.146)	(0.054)	(1.36)

Tabela 2- Médias empíricas e o Erro quadrático médio (entre parênteses) das estimativas para os parâmetros α e λ, para valores de T, utilizando os métodos de Yule-Walker, Squared Difference e Kendall. Dados simulados com $\lambda = 1$ e os valores correspondentes de α e T, com outliers de magnitude de 10.

		$\alpha = 0.2$		$\alpha = 0.5$		$\alpha = 0.8$	
Método	T	α	λ	α	λ	α	λ
Yule-Walker	100	0.123	1.19	0.340	1.39	0.633	1.87
		(0.016)	(0.08)	(0.046)	(0.290)	(0.047)	(1.30)
	300	0.114	1.20	0.338	1.39	0.656	1.76
		(0.012)	(0.056)	(0.034)	(0.203)	(0.028)	(0.763)
	500	0.113	1.20	0.335	1.40	0.661	1.73
		(0.010)	(0.050)	(0.032)	(0.186)	(0.023)	(0.646)
Squared Difference	100	0.076	1.98	0.217	1.99	0.609	1.99
		(0.031)	(2.01)	(0.134)	(2.06)	(0.073)	(2.02)
	300	0.013	1.99	0.137	1.98	0.609	1.99
		(0.037)	(1.32)	(0.156)	(1.32)	(0.049)	(1.34)
	500	0.004	1.99	0.111	1.99	0.610	1.99
		(0.039)	(1.19)	(0.168)	(1.19)	(0.043)	(1.18)
Kendall	100	0.102	1.21	0.282	1.50	0.511	2.49
		(0.013)	(0.072)	(0.051)	(0.302)	(0.089)	(2.42)
	300	0.107	1.20	0.297	1.48	0.541	2.34
		(0.010)	(0.051)	(0.043)	(0.242)	(0.069)	(1.87)
	500	0.109	1.20	0.300	1.47	0.546	2.31
		(0.009)	(0.046)	(0.041)	(0.230)	(0.065)	(1.77)

Observando a Tabela 3, queremos que os valores obtidos sejam próximos de zero, visto que se o valor real da série temporal simulada e o valor predito forem iguais, então a diferença entre eles seria zero. Desse modo, percebe-se que os valores obtidos com $\alpha = 0.2$ não estão próximos de zero, enquanto os outros estão.

Tabela 3- Médias empíricas e o Erro quadrático médio (entre parênteses) das diferenças entre o valor real e o valor previsto dos dados simulados com e sem outliers. Os valores preditos foram obtidos através dos parâmetros estimados α e λ pelos métodos Yule-Walker, Squared Difference e Kendall, com sua distribuição acumulada condicional.

		Yule-Walker		Squared Difference		Kendal	
T	α	Sem	Com	Sem	Com	Sem	Com
100	0.2	0.190	0.165	0.179	-0.664	0.226	0.177
		(1.30)	(1.34)	(1.33)	(2.63)	(1.31)	(1.33)
	0.5	0.125	0.048	0.118	-0.294	0.124	0.0473
		(1.60)	(1.78)	(1.60)	(2.460	(1.67)	(1.80)
	0.8	0.0694	0.001	0.0671	-0.016	0.0978	0.047
		(1.94)	(2.46)	(1.92)	(2.63)	(2.23)	(2.63)
300	0.2	0.210	0.195	0.201	-0.565	0.258	0.202
		(1.31)	(1.28)	(1.31)	(1.97)	(1.33)	(1.28)
	0.5	0.184	0.0375	0.178	-0.120	0.142	0.0192
		(1.55)	(1.71)	(1.55)	(2.02)	(1.61)	(1.72)
	0.8	0.0853	-0.0256	0.0818	-0.0177	0.110	-0.0009
		(1.88)	(2.46)	(1.87)	(2.52)	(2.15)	(2.57)
500	0.2	0.170	0.213	0.166	-0.563	0.221	0.215
		(1.25)	(1.34)	(1.26)	(1.91)	(1.27)	(1.35)
	0.5	0.216	0.0347	0.212	-0.0696	0.150	0.0132
		(1.61)	(1.77)	(1.61)	(2.00)	(1.67)	(1.77)
	0.8	0.0753	-0.029	0.0794	-0.0222	0.112	0.0084
		(1.88)	(2.43)	(1.88)	(2.50)	(2.15)	(2.52)

CONCLUSÕES

Este projeto de iniciação científica estudou o modelo INAR, restringido ao INAR(1), para modelagem séries temporais para dados discretos, tanto na estimação de parâmetros quanto para a previsão de valores futuros. Os métodos de previsão e estimação foram implementados em R e serão disponibilizados em um repositório no GitHub.

Dos resultados, percebe-se que o método de previsão é viável utilizando os métodos de estimação apresentados na pesquisa, não tendo uma diferença significativa nos resultados entre Yule-Walker e Squared Difference. Contudo, se houver outliers ao longo dos dados na série temporal analisada, é recomendável utilizar o método de Kendall para estimar os parâmetros e fazer a previsão.

BIBLIOGRAFIA

Alzaid, A. A., & Al-Osh, M. A. (1987). First-order integer-valued autoregressive (INAR (1)) process. *Journal of Time Series Analysis, 8*(3), 261-275. doi:10.1111/j.1467-9892.1987.tb00438.x

Bourguignon, M., & Vasconcellos, K. L. (2015). Improved estimation for Poisson INAR(1) models. *Journal of Statistical Computation and Simulation, 85*. doi:10.1080/00949655.2014.930862

Bourguignon, M., & Vasconcellos, K. L. (2018). The effects of additive outliers in {INAR (1)} process and robust estimation. *Statistical Theory and Related Fields, 2*, 206-214. doi:10.1080/24754269.2018.1520018

Freeland, R. K., & McCabe, B. P. (2004). Forecasting discrete valued low count time series. *International Journal of Forecasting, 20*, 427-434. doi:10.1016/S0169-2070(03)00014-1

Harn, K., & Steutel, F. W. (1979). Discrete Analogues of Self-Decomposability and Stability. *The Annals of Probability, 7*, 893-899. Fonte: http://www.jstor.org/stable/2243313

Jin-Guan, D., & Li, Y. (1991). The integer-valued autoregressive (INAR(p)) model. *Journal of Time Series Analysis, 12*, 129-142. doi:10.1111/j.1467-9892.1991.tb00073.x

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*. Fonte: https://doi.org/10.2307/2332226

Pavlopoulos, H., & Karlis, D. (2008). INAR (1) modeling of overdispersed count series with an environmental application. *Environmetrics, 19*, 369-393. doi:10.1002/env.883

Silva, M. E. (2015). *Modelling time series of counts: An inar approach.*

Turkman, K. F., Scotto, M. G., & Bermudez, P. d. (2014). Models for Integer-Valued Time Series. Em *Non-Linear Time Series: Extreme Events and Integer Value Problems.* Springer International Publishing. doi:10.1007/978-3-319-07028-5_5