Subamostragem em processos autoregressivos de valores inteiros

Sara Morgado Nunes
Departamento de Matemática Aplicada - FCUP e Escola Superior de Gestão - IPCB (sara@esg.ipcb.pt)

Maria Eduarda Silva
Departamento de Matemática Aplicada - FCUP (mesilva@fc.up.pt)

Resumo: O modelo AutoRegressivo de valor INteiro, INAR, foi proposto na literatura para modelar séries de contagem. Vários métodos de estimação para os parâmetros do modelo têm sido propostos. No entanto, as propriedades assimptóticas dos estimadores dos parâmetros, nomeadamente a variância, são difíceis de obter, impossibilitando a construção de intervalos de confiança. Neste trabalho, estuda-se a técnica de subamostragem, aplicando-a à construção de intervalos de confiança para os parâmetros do modelo INAR(1).

Palavras–chave: subamostragem, série temporal, processo INAR, intervalo de confiança

Abstract: The INteger-valued AutoRegressive models, INAR, have been proposed in the literature to model count series. The estimation of the parameters of this model can be acomplished using several approaches. However, the properties of the estimators, namely the variance, are difficult to obtain, thus impairing the construction of confidence intervals for the parameters. Here, subsampling is used to obtain confidence intervals for the parameter of a INAR(1) model.

Keywords: subsampling, time series, INAR process, confidence interval

1 Introdução

Tem-se registado recentemente um interesse crescente da comunidade científica pelos modelos INAR pois muitas das séries que se observam são séries de valores inteiros não negativos e, em particular, séries de contagem.

Neste caso, os modelos a que habitualmente se recorre na modelação de séries temporais, tanto lineares como não lineares, revelam-se inadequados, uma vez que do produto de uma constante real por uma variável aleatória de valor inteiro resulta uma variável aleatória real. Com o objectivo de solucionar este problema, McKenzie (1986,1988) e Al-Osh e Alzaid (1987) recorreram à operação *thinning* binomial definida por Steutel e Van Harn (1979), substituindo a operação de multiplicação usual e propondo os modelos INAR(1), que passamos a definir.

Seja $\{X_t\}$ um processo estocástico de valores inteiros não negativos, $a \in [0,1]$ e $*$ a operação *thinning* binomial assim definida [Steutel e Van Harn (1979)]:

$$a * X = \sum_{k=1}^{X} Y_k \tag{1}$$

onde $\{Y_k\}$, dita série de contagem, é uma sequência de variáveis aleatórias independentes e identicamente distribuídas (i.i.d.), independentes de X, tais que $P(Y_k = 1) = 1 - P(Y_k = 0) = a$. Então $\{X_t\}$, diz-se um processo INAR(1) se satisfaz a seguinte equação

$$X_t = a * X_{t-1} + e_t \tag{2}$$

onde $a \in]0,1[$, $t \in \{0, \pm 1, \pm 2, ...\}$ e $\{e_t\} \in \mathbb{N}_0$ é uma sequência de variáveis aleatórias i.i.d., com média μ_e e variância σ_e^2.

Se a sequência de variáveis i.i.d., $\{e_t\}$, tem distribuição de Poisson, o processo é dito INAR de Poisson, sendo os processos INAR(1) assim definidos sempre estacionários.

Mostra-se (Alzaid e Al-Osh (1987); Silva e Oliveira (2000)) que a estrutura de correlação dos processos INAR(p) é semelhante à dos processos AR(p), podendo, a sua função de densidade espectral, $f(\omega)$, escrever-se como

$$f(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} R(k)\, e^{-i\omega k} = \frac{\mu_e(1+a)}{2\pi(1 - 2a\cos\omega + a^2)}, \; -\pi \leq \omega \leq \pi. \tag{3}$$

O problema da estimação dos parâmetros a, μ_e e σ_e^2 do modelo tem sido considerada por diversos autores. Al-Osh e Alzaid (1987) propuseram a estimação de Yule-Walker, máxima verosimilhança e o método dos mínimos quadrados condicionais. No domínio da frequência, Silva e Oliveira (2000) propuseram um método de estimação baseado na minimização do critério de Whittle.

No entanto, apesar da distribuição assimptótica dos estimadores de Yule-Walker, mínimos quadrados condicionais e de máxima verosimilhança ser normal, expressões para a variância assimptótica são difíceis de obter. Relativamente aos estimadores obtidos pela minimização do critério de Whittle, a sua variância assimptótica depende dos cumulantes de 4a ordem do processo. Assim, não é possível a construção de intervalos de confiança para o parâmetro a, pelo que a inferência estatística sobre o modelo estimado fica, de certo modo, comprometida.

Foi, precisamente, numa tentativa de medir a precisão de uma estatística $\hat{\theta}$ enquanto estimador de θ e de estimar o seu erro padrão que surgiram os métodos de reamostragem como o bootstrap [Efron (1979)] e de subamostragem como o jackknife [Tukey (1958)]. Estes métodos foram primeiramente propostos para observações de variáveis i.i.d., tendo sido desenvolvidas, posteriormente, metodologias que permitem a sua extensão a séries temporais. É neste contexto que Politis e Romano (1992, 1994) propõem a construção de intervalos de confiança para um parâmetro θ com base no método de subamostragem cujo princípio

básico é o cálculo da estatística de interesse em subconjuntos da amostra original, de modo a obter uma aproximação da distribuição amostral de $\hat{\theta}$ e fazer depois inferências sobre θ.

Neste trabalho aplica-se a metodologia da subamostragem a séries geradas por modelos INAR(1). Como atrás ficou dito, o cálculo da variância associada a um determinado conjunto de estimativas do parâmetro do modelo (a) é extremamente complicado, o que faz com que não se disponham de critérios para verificar se o parâmetro estimado é ou não significativo. Foi justamente a existência desta dificuldade que motivou a aplicação da metodologia da subamostragem a estes modelos. Assim, na secção 2 são tecidas algumas considerações acerca da metodologia da subamostragem e, na secção 3, são apresentados os resultados de um estudo de simulação levado a efeito com o objectivo de construir intervalos de confiança para o parâmetro de processos INAR(1).

2 Subamostragem em Séries Temporais

Sejam $\{X_1, X_2, \ldots, X_n\}$ n observações de uma série temporal estacionária e que satisfaz uma condição de dependência fraca, com valores num espaço amostral $\mathcal{S}$ e seja $\mathcal{P}$ uma medida de probabilidade. Suponha-se que se opta por uma estatística $\hat{\theta}$ para estimar um determinado parâmetro $\theta = \theta(\mathcal{P}) \in \mathbb{R}$. O objectivo do método de subamostragem introduzido por Politis e Romano (1992, 1994) é a construção de intervalos ou regiões de confiança para θ através da aproximação da distribuição amostral de $(\hat{\theta}_n - \theta(\mathcal{P}))$ pela distribuição empírica dos valores da estatística calculada em subamostras formadas por blocos de observações consecutivas de dimensão $b < n$, sendo a primeira $\{X_1, X_2, \ldots, X_b\}$ e a última $\{X_{n-b+1}, X_{n-b+2}, \ldots, X_n\}$. Note-se que existem $q = n - b + 1$ blocos destes.

Os valores resultantes do cálculo da estatística em cada subamostra são convenientemente normalizados de forma a aproximarem a verdadeira distribuição amostral em causa. Deste modo, a subamostragem constitui um método muito geral para a construção de intervalos de confiança de primeira ordem assimptoticamente válidos pois, se b for tal que $b/n \to 0$ e $b \to \infty$ quando $n \to \infty$, o método é válido sempre que a estatística original, convenientemente normalizada, apresenta uma distribuição limite sob o verdadeiro modelo.

Seja $\hat{\theta}_{n,b,t} = \hat{\theta}_b(X_t, \ldots, X_{t+b-1})$ o estimador de $\theta(\mathcal{P})$ baseado na subamostra $\{X_t, \ldots, X_{t+b-1}\}$ (note-se que, de acordo com esta notação, $\hat{\theta}_n = \hat{\theta}_{n,n,1}$). Seja $J_{b,t}(\mathcal{P})$ a distribuição amostral de $\tau_b(\hat{\theta}_{n,b,t} - \theta(\mathcal{P}))$, onde τ_b é uma constante de normalização apropriada. Defina-se também a correspondente função distribuição acumulada como $J_{b,t}(x, \mathcal{P}) = \text{Prob}_{\mathcal{P}}\{\tau_b(\hat{\theta}_{n,b,t} - \theta(\mathcal{P})) \leq x\}$.

Por conveniência, denote-se por $J_n(\mathcal{P}) = J_{n,1}(\mathcal{P})$ a distribuição amostral de $\tau_n(\hat{\theta}_n - \theta(\mathcal{P}))$ e por $J_n(\cdot, \mathcal{P})$ a função distribuição acumulada correspondente.

A aproximação subamostragem a $J_n(x,\mathcal{P})$ é então definida por:

$$L_{n,b}(x) = \frac{1}{n-b+1} \sum_{t=1}^{n-b+1} I\{\tau_b(\hat{\theta}_{n,b,t} - \hat{\theta}_n) \leq x\} \tag{4}$$

onde $I\{E\}$ é a função indicatriz do evento $E = \{\tau_b(\hat{\theta}_{n,b,t} - \hat{\theta}_n) \leq x\}$.

Para cada t, $\{X_t, \ldots, X_{t+b-1}\}$ é uma subamostra de tamanho b do verdadeiro modelo $\mathcal{P}$. Logo, a distribuição exacta de $\tau_b(\hat{\theta}_{n,b,t} - \theta(\mathcal{P}))$ é $J_b(\mathcal{P})$. A estacionaridade implica que a distribuição empírica dos $n-b+1$ valores de $\tau_b(\hat{\theta}_{n,b,t} - \theta(\mathcal{P}))$ constitui uma boa aproximação a $J_n(\mathcal{P})$. Assim, sendo $\theta(\mathcal{P})$ desconhecido, podemos substituir $\theta(\mathcal{P})$ por $\hat{\theta}_n$ uma vez que $\tau_b(\hat{\theta}_n - \theta(\mathcal{P}))$ é de ordem τ_b/τ_n em probabilidade, assumindo-se que $\tau_b/\tau_n \to 0$.

As condições sob as quais o método de subamostragem conduz a resultados assimptoticamente válidos para estatísticas gerais, funcionando sob condições mínimas, foram obtidas por Politis e Romano (1994) e estão estabelecidas no seguinte teorema:

Teorema 1 (Politis e Romano (1994)). *Notemos a sequência α – mixing correspondente a $\{X_t\}$ por $\alpha_X(\cdot)$.*

Assumamos a Hipótese A e que $\tau_b/\tau_n \to 0$, $b/n \to 0$ e $b \to \infty$ quando $n \to \infty$. Assumamos também que $\alpha_X(m) \to 0$ quando $m \to \infty$.

1. *Se x é um ponto de continuidade de $J(\cdot,\mathcal{P})$, então $L_{n,b}(x) \to J(x,\mathcal{P})$ em probabilidade.*

2. *Se $J(\cdot,\mathcal{P})$ é contínua, então $sup_x|L_{n,b}(x) - J(x,\mathcal{P})| \to 0$ em probabilidade.*

3. *Para $\alpha \in [0,1]$, seja $c_{n,b}(1-\alpha) = inf\{x : L_{n,b}(x) \geq 1-\alpha\}$.*

Correspondentemente, definamos $c(1-\alpha,\mathcal{P}) = inf\{x : J(x,\mathcal{P}) \geq 1-\alpha\}$. Se $J(\cdot,\mathcal{P})$ é contínua em $c(1-\alpha,\mathcal{P})$, então $Prob_{\mathcal{P}}\{\tau_n[\hat{\theta}_n - \theta(\mathcal{P})] \leq c_{n,b}(1-\alpha)\} \to 1-\alpha$ quando $n \to \infty$. Então, a probabilidade de cobertura assimptótica sob $\mathcal{P}$ do intervalo $I_1 = [\hat{\theta}_n - \tau_n^{-1}c_{n,b}(1-\alpha), \infty[$ é $1-\alpha$.

Um intervalo de confiança bilateral de caudas iguais pode ser obtido através da intersecção de dois intervalos unilaterais. Assim, o intervalo bilateral análogo a I_1 é

$$I_{ET} = \left[\hat{\theta}_n - \tau_n^{-1}c_{n,b}(1-\frac{\alpha}{2}), \hat{\theta}_n - \tau_n^{-1}c_{n,b}(\frac{\alpha}{2})\right]. \tag{5}$$

I_{ET} diz-se um intervalo de caudas iguais por ter probabilidade aproximadamente igual em cada cauda:

$$\text{Prob}_{\mathcal{P}}\left\{\theta < \hat{\theta}_n - \tau_n^{-1}c_n(1-\frac{\alpha}{2})\right\} \approx \text{Prob}_{\mathcal{P}}\left\{\theta > \hat{\theta}_n - \tau_n^{-1}c_n(\frac{\alpha}{2})\right\} \approx \frac{\alpha}{2}.$$

Como aproximação alternativa, também podem ser construídos intervalos de confiança simétricos bilaterais. Para isso estimamos a função distribuição bilateral

$$J_{n,|\cdot|}(x,\mathcal{P}) = \text{Prob}_{\mathcal{P}}\{\tau_n|\hat{\theta}_n - \theta(\mathcal{P})| \leq x\}.$$

A aproximação por subamostragem a $J_{n,|\cdot|}(x,\mathcal{P})$ é então definida por

$$L_{n,b,|\cdot|}(x) = \frac{1}{n-b+1}\sum_{t=1}^{n-b+1} I\{\tau_b|\hat{\theta}_{n,b,t} - \hat{\theta}_n| \leq x\}. \quad (6)$$

Denotando um quantil $(1-\alpha)$ de $L_{n,b,|\cdot|}$ por $c_{n,b,|\cdot|}(1-\alpha)$, o intervalo subamostragem simétrico é então dado por

$$I_{sym} = [\hat{\theta}_n - \tau_n^{-1}c_{n,b,|\cdot|}(1-\alpha), \hat{\theta}_n + \tau_n^{-1}c_{n,b,|\cdot|}(1-\alpha)]. \quad (7)$$

Intervalos de confiança simétricos não constituem necessariamente uma escolha superior pois a assimetria de um intervalo de caudas iguais pode conter informação útil acerca da localização do verdadeiro parâmetro e sobre a assimetria da distribuição amostral do estimador.

3 Estudo de Simulação

Neste estudo de simulação geraram-se amostras de 256 observações do processo INAR(1), definido em (2), onde a sequência de inovações tem distribuição $Po(1)$ e o parâmetro a do modelo é o parâmetro de interesse. Como possíveis critérios de estimação do parâmetro em causa dispõe-se dos estimadores de Yule-Walker, máxima verosimilhança, mínimos quadrados condicionais e critério de Whittle, tendo-se recorrido, neste estudo, ao primeiro e ao último. Assim, nas amostras geradas, o parâmetro a tomou sucessivamente os valores 0.2, 0.5, 0.8, 0.95 e 0.99 e a variância das inovações $\sigma_e^2 = 1, 3, 5$. Às amostras geradas aplicou-se a subamostragem usando diferentes tamanhos de bloco ($b = 4$, $b = 8$, $b = 16$, $b = 32$, $b = 64$ e $b = 128$). Para cada situação construiu-se, para o parâmetro do modelo, um intervalo de confiança a 95%, de caudas iguais I_{ET}, (5), e um intervalo de confiança a 95% simétrico I_{sym}, (7). Repetiu-se este procedimento 500 vezes e registou-se, em cada simulação, o número de intervalos de confiança que continham o verdadeiro parâmetro com o objectivo de estimar a probabilidade de cobertura.

As probabilidades de cobertura estimadas para intervalos de confiança a 95% por subamostragem de caudas iguais - I_{ET} - e simétricos - I_{sym} - para o parâmetro do modelo - a, usando diversos tamanhos de bloco - b - recorrendo aos estimadores de Yule-Walker e de Whittle, apresentam-se na tabela 1. Na subamostragem com estimação pelo critério de Whittle não foram usadas amostras de tamanho menor que 16, por não ser viável a aplicação deste critério em amostras de dimensão tão pequena. Apresenta-se somente o resultado das

		b									
		4	8	16		32		64		128	
a		YW	YW	YW	W	YW	W	YW	W	YW	W
0	I_{ET}	0.62	0.85	0.86	0.91	0.87	0.90	0.78	0.81	0.58	0.62
	I_{sym}	0.40	0.71	0.76	0.91	0.78	0.86	0.70	0.77	0.50	0.58
0.2	I_{ET}	0.57	0.78	0.86	0.93	0.86	0.87	0.80	0.81	0.60	0.59
	I_{sym}	0.10	0.54	0.68	0.86	0.72	0.81	0.13	0.74	0.52	0.56
0.5	I_{ET}	0.42	0.67	0.81	0.96	0.84	0.90	0.78	0.81	0.60	0.62
	I_{sym}	0	0.16	0.46	0.91	0.59	0.77	0.62	0.70	0.48	0.52
0.8	I_{ET}	0.16	0.46	0.63	0.91	0.80	0.97	0.80	0.88	0.60	0.62
	I_{sym}	0	0	0.03	0.82	0.33	0.86	0.42	0.70	0.35	0.47
0.95	I_{ET}	0.03	0.25	0.49	0.68	0.66	0.76	0.78	0.76	0.67	0.62
	I_{sym}	0	0	0	0.37	0	0.54	0.06	0.73	0.17	0.66
0.99	I_{ET}	0.01	0.12	0.31	0.61	0.59	0.67	0.76	0.84	0.76	0.67
	I_{sym}	0	0	0	0.15	0	0.34	0	0.78	0	0.75

Tabela 1: Probabilidade de cobertura estimada para intervalos de confiança a 95% por subamostragem de caudas iguais - I_{ET} - e simétricos - I_{sym} - para o parâmetro do modelo - a, usando diversos tamanhos de bloco - b - e recorrendo aos estimadores de Yule-Walker - YW - e de Whittle - W.

simulações feitas com $\sigma_e^2 = 3$ por não ter sido detectada a existência de uma relação directa entre σ_e^2 e a probabilidade de cobertura estimada.

Da análise das tabela retiram-se as seguintes conclusões:

- os estimadores obtidos pelo critério de Whittle produzem intervalos de confiança para o parâmetro com uma maior probabilidade de cobertura;
- os intervalos de confiança de caudas iguais apresentam, neste caso, uma probabilidade de cobertura superior à dos intervalos de confiança simétricos;
- para valores do parâmetro do modelo próximos de 1 ($a = 0.8$, $a = 0.95$ e $a = 0.99$), os intervalos de confiança têm tendência a apresentar probabilidades de cobertura inferiores;
- o tamanho de bloco (b) que conduz a melhores resultados aumenta à medida que o valor do parâmetro do modelo aumenta. Por exemplo para $a = 0.2$ parece ser $b = 16$ que conduz a intervalos de confiança com melhor cobertura, enquanto que para $a = 0.8$ é $b = 64$. Assim, o tamanho de bloco óptimo depende, claramente, do valor do parâmetro pois, tamanhos de blocos maiores permitem captar melhor estruturas de dependência fortes.

Para $a = 0$, caso em que as observações são i.i.d., as conclusões não diferem significativamente do que atrás ficou dito.

Na figura 1 apresentam-se, para $a = 0.5$ e $\sigma_e^2 = 3$, os limites dos 500 intervalos de confiança de caudas iguais usando: **(a)** o método usado na estimação do parâmetro foi o critério de Whittle com subamostras de tamanho 16, tendo sido a probabilidade de cobertura estimada 0.96 e **(b)** os estimadores de Yule-Walker, com subamostras de tamanho 32, tendo sido a probabilidade de cobertura estimada 0.84.

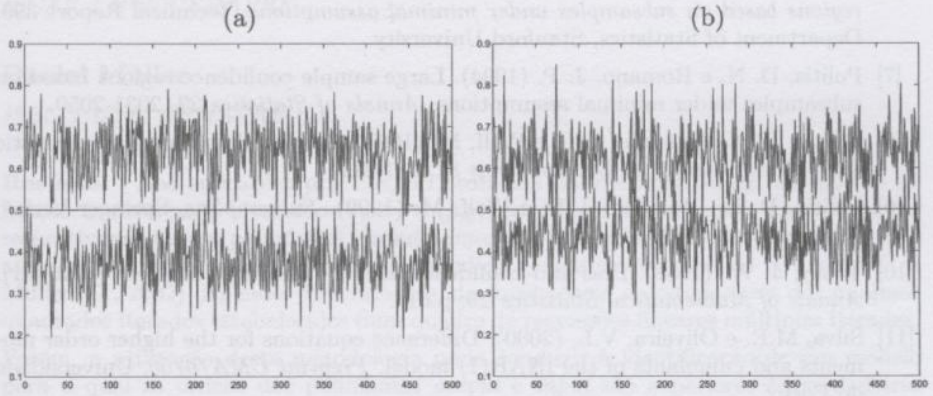

Figura 1: Limites inferiores e superiores dos 500 intervalos de confiança de caudas iguais (I_{ET}) a 95% para $a = 0.5$, $\sigma_e^2 = 3$, usando estimadores de: **(a) Whittle,** com $b = 16$ e **(b) Yule-Walker,** com $b = 32$.

4 Comentários Finais

Constata-se que a metodologia da subamostragem se aplica sob condições muito gerais, constituindo um instrumento na construção de intervalos de confiança para parâmetros cuja expressão para a variância é difícil de obter. Neste trabalho aplica-se a subamostragem a séries temporais geradas por processos INAR(1), com o objectivo de construir intervalos de confiança para o parâmetro do modelo. Este estudo pretende ser um primeiro passo no desenvolvimento de uma metodologia para a avaliação da qualidade de ajuste de modelos INAR.

Bibliografia

[1] Al-Osh, M.A. e Alzaid, A.A.(1987). First-order integer-valued autoregressive (INAR(1)) process. *J. Time Series Anal.*, Vol. 8, pp. 261-75.

[2] Bertail, P., Politis, D. N. e Romano, J. P. (1999). On subsampling estimators with unknown rate of convergence. *J. American Stat. Assoc.* 94, 569-579.

[3] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* 7, 1-26.

[4] McKenzie, E. (1986). Autoregressive moving-average process with negative-binomial and geometric marginal distributions. *Adv. Appl. Probab.*, Vol. 18, pp. 679-705.

[5] McKenzie, E. (1988). Some ARMA models for dependent sequences of Poisson counts. *Adv. Appl. Probab.*, Vol. 20, pp. 822-35.

[6] Politis, D. N. e Romano, J. P. (1992). *A general theory for large sample confidence regions based on subsamples under minimal assumptions.* Technical Report 399, Department of Statistics, Stanford University.

[7] Politis, D. N. e Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Annals of Statistics* 22, 2031-2050.

[8] Politis, D. N., Romano, J. P. e Wolf, M. (1997). Subsampling for heteroskedastic time series. *Journal of Econometrics* 81, 281-317.

[9] Politis, D. N., Romano, J. P. e Wolf, M. (1999). *Subsampling.* Springer-Verlag, Nova Iorque.

[10] Tukey, J. W. (1958). Bias and confidence in not quite large samples (abstract). *Annals of Mathematical Statistics* 29, p.614.

[11] Silva, M.E. e Oliveira, V.L. (2000). Difference equations for the higher order moments and cumulants of the INAR(1) model. *Preprint CMA/6/00*, Universidade do Porto.

[12] Steutel, F.W. e Van Harn, K. (1979). Discrete analogues of self-decomposability and stability. *Ann. Probab.*, Vol. 7, pp. 893-99.