Forecasting interval for the INAR(p) process using sieve bootstrap

Hee-Young Kim and Yousung Park *

Abstract

Recently, as a result of the growing interest in modelling stationary processes with discrete marginal distributions, several models for integer valued time series have been proposed in the literature. One of theses models is the integer-valued autoregressive(INAR) models. However, when modelling with integer-valued autoregressive processes, there is not yet distributional properties of forecasts, since INAR process contain an accrued level of complexity in using the Steutal and Van Harn(1979) thinning operator " $\circ$ ".

In this study, a manageable expression for the asymptotic mean square error of predicting more than one-step ahead from an estimated poisson INAR(1) model is derived. And, we present a bootstrap methods developed for the calculation of forecast interval limits of INAR(p) model. Extensive finite sample Monte Carlo experiments are carried out to compare the performance of the several bootstrap procedures.

Keywords : Stationary process, Integer valued time series, Sieve bootstrap.

1 Introduction

When studying a time series, one of the main goals is the estimation of the forecast intervals based on an observed sample path of the process. That is, providing information about the distribution of the variable X_{T+h} conditional on a realization of the past variables $\mathbf{X}_T = \{X_1, \cdots, X_T\}$ is the main objective. In particular, the aim is to construct prediction intervals $I(\mathbf{X}_T) = \{L((X_T), U(\mathbf{X}_T)\}$ designed to capture the future value of X_{T+k} with a fixed probability, the nominal coverage. But INAR(p) process

$$X_t = \alpha_1 \circ X_{t-1} + \alpha_2 \circ X_{t-2} + \cdots + \alpha_p \circ X_{t-p} + \epsilon_t$$

has Steutal and Van Harn operator "$\circ$", it is not easy task to obtain distributional properties of forecasts. Recent work by Freeland and McCabe(2004) focus on the coherent forecast, in the sense of producing only integer forecasts of count variable. They propose point forecasts in the context of the poisson INAR(1) model, based on the integer-valued median of the forecast distribution. Estimation uncertainty is accommodated by producing a maximum likelihood-based estimate of the forecast distribution and constructing confidence intervals around the estimated probabilities. And McCabe and Marin(2005) presents a Bayesian methodology for producing coherent forecasts

[*] Hee-Young Kim is an research associate professor in the Institute of statistics, Korea University, Korea; Yousung Park is a professor, Department of Statistics, Korea University

of count time series. Both research are based on estimates of the h-step ahead predictive mass functions.

In contrast, in this paper, we get the formula of the mean square error of the predictor error conditional on $\{X_T, \cdots, X_1\}$ for the poisson INAR(1) process. However, we obtained explicit formula only for poisson INAR(1) process in the present. Of course, poisson based prediction intervals fail if the data are not poisson distributed.

Therefore, we proposes an alternative method that avoid theses problems. The bootstrap provides an estimate of the conditional distribution of X_{T+h}.

The paper is organized as follows. In Section 2, INAR(p) processes are described in some detail. We analyze, in section 3, the unconditional mean squared error of the predictor of poisson INAR(1) model. Section 4 introduces the sieve bootstrap for estimating forecasts intervals for INAR(p) process. Section 5 presents a Monte carlo study comparing the finite sample properties of the sieve bootstrap. We show that the result of sieve bootstrap for INAR(p) are not the similar for AR(p) process.

2 Time series models for counts

The INAR(1) processes is defined by the difference equation

$$X_t = \alpha_1 \circ X_{t-1} + \epsilon_t \tag{1}$$

with the state spce of the process being $N_0 = N \bigcup \{0\}$. It is assumed that $\alpha_1 \in [0, 1)$ and ϵ_t is a sequence of i.i.d non-negative integer-valued random variables with mean $E(\epsilon_t) = \mu_\epsilon, Var(\epsilon_t) = \sigma_\epsilon^2$. The process $\{X_t\}$ satisfying (1) is a second- order staionary if $0 \leq \alpha_1 < 1$ (Al-Osh and Alzaid, 1987, Du and Li, 1991). The probability generating function of the random variable $X_t, \phi_{X_t}(s)$, is given by Alzaid and Al-Osh(1988), $\phi_{X_t}(s) = \phi_{X_0}(1 - \alpha^t + \alpha^t s) \prod_{k=0}^{t-1} \phi_e(1 - \alpha^k - \alpha^k s)$, $|s| \leq 1$, where $\phi_e(s)$ is the probability generating function of ϵ_1. Since $E(\epsilon_1) < \infty$, the limit $\lim_{t \to \infty} \phi_{X_t}(s)$ exists and is the probability generating function of some random variable X, for which

$$\phi_X(s) = \phi_e(s) \phi_X(1 - \alpha + \alpha s) \tag{2}$$

(Alzaid and Al-Osh, 1988). Equation (2) is related to the definition of a self-decomposable distribution on a set of non-negative integers as introduced by Steutel and Van Harn(1979). In fact, assuming that the INAR(1) process is stationary, the probability generating function of X_t satisfies (2). Therefore, one can choose any member of the class of discrete self-decomposable distributions as the marginal of a stationary INAR(1) process.

The INAR(p) process $\{X_t\}$, a seemingly natural extension of the INAR(1) process, is defined in the usual manner

$$X_t = \alpha_1 \circ X_{t-1} + \alpha_2 \circ X_{t-2} + \cdots + \alpha_p \circ X_{t-p} + \epsilon_t \tag{3}$$

where

(a) $\{\epsilon_t\}$ is a sequence of i.i.d non-negative integer-valued random variables, with $E(\epsilon_t) = \mu_\epsilon, Var(\epsilon_t) = \sigma_\epsilon^2, E(\epsilon_t^3) < \infty$

(b) All counting series of $\alpha_i \circ X_{t-i}, i = 1, \cdots, p$, $\{Y_{i,k}\}, k = 1, \cdots, X_{t-i}$ are mutually independent, and independent of $\{\epsilon_t\}$

(c) $0 < \alpha_i < 1, i = 1, \cdots, p$.

To ensure the stationarity of the process, Du and Li(1991) proved that all the roots of $\lambda^p - \alpha_1\lambda^{p-1} - \cdots \alpha_p\lambda - \alpha_p = 0$ are inside the unit circle is the stationarity condition. The mean of the process satisfies $E(X_t) = \mu_\epsilon/(1 - \sum_{i=1}^{p} \alpha_i)$ and the autocorrelation function of the process is the same as that of an AR(p).

3 Properties of predictor for INAR(1) process

In this section, we derive the prediction mean squared error of predictor for poisson INAR(1) process. In the INAR(1) process, the Poisson distribution plays a role similar to that of the Gaussian distribution in the AR(1) process. It has the property that if the innovation sequence $\{\epsilon\}$ and the initial distribution are poisson, the marginal distribution of $\{X_t\}$ is also poisson. The most common procedure for constructing forecasts in time series models is to use conditional expectations. The reason is that this technique will yield forecasts with minimum mean square error. That is, given $\theta = (\theta_1, \theta_2)' = (\alpha, \mu_\epsilon)'$ and $\mathbf{X} = (x_1, \cdots, x_T)'$, the minimum MSE forecast of X_{T+h} is $\tilde{X}_{T+h} = E(X_{T+h}|\mathcal{F}_T)$.

Theorem 3.1. *Let $\{X_t\}$ be a stationary process satisfying (1) where $\{\epsilon_t\}$ are i.i.d random variables, independent of $\{X_t\}$, with poisson distribution with parameter λ. Then the PMSE of $\tilde{X}_{T+h} = E(X_T|\mathcal{F}_T)$ is given by*

$$E[X_{T+h} - \tilde{X}_{T+h}]^2 = \frac{\lambda(1-\alpha^{2h})}{1-\alpha}$$

Typically, the $\theta = (\theta_1, \theta_2)' = (\alpha, \lambda)'$ are not known, these parameters are estimated. The least squares estimator of the θ_i's, say $\hat{\theta}_i$'s is often used and this leads to parameter estimates which are asymptotically normal.

The estimated predictor is computed by replacing θ by $\hat{\theta}$. The predictor of X_{T+h} with estimated coefficients $\hat{\theta}$, say $\hat{X}_{T+h}$ is given by

$$\hat{X}_{T+h} = \hat{\alpha}^h X_T + \hat{\lambda}\frac{1-\hat{\alpha}^h}{1-\hat{\alpha}} \tag{4}$$

By Taylor expansion of (4) around $\hat{\theta} = \theta$,

$$\begin{aligned}\hat{X}_{T+h} &= \alpha^h X_T + \lambda\frac{1-\alpha^h}{1-\alpha} \\ &+ \left(h\alpha^{h-1}X_T + \lambda\frac{-h\alpha^{h-1} + h\alpha^h + 1 - \alpha^h}{(1-\alpha)^2}\right)(\hat{\alpha} - \alpha) \\ &+ \frac{1-\alpha^h}{1-\alpha}(\hat{\lambda} - \lambda) + \textit{remainder terms}\end{aligned}$$

We conventionally assume that statistical independent between the processes in estimation and prediction. This independence assumptions are made in several papers about in usual pth-order AR model(Bloomfield, 1972; Bhansali, 1974; Schmidt, 1974;Yamamoto, 1976)

Theorem 3.2. *Let $\{X_t\}$ be a stationary process satisfying (1) where $\{\epsilon_t\}$ are i.i.d random variables, independent of $\{X_t\}$, with poisson distribution with parameter λ. Then the asymptotic mean square error of $\hat{X}_{T+h}$ is obtained as*

$$E(\hat{X}_{T+h} - X_{T+h})^2$$

$$\begin{aligned}
&= \frac{1}{T}E\Big(h^2\alpha^{2(h-1)}X_T^2 + 2h\alpha^{h-1}AX_T + A^2\Big)\frac{1-\alpha}{\lambda}(\alpha\lambda + \alpha - \alpha^2 + \lambda) \\
&+ \frac{1}{T}\frac{(1-\alpha^h)^2\{\alpha(\lambda^4 - \lambda^3) + (\lambda^4 + \lambda^3)\}}{(1-\alpha)^3\lambda^2} \\
&- 2\frac{1}{T}\frac{1-\alpha^h}{1-\alpha}E(h\alpha^{h-1}X_T + A)\lambda^2(1+\alpha) \\
&+ \lambda\frac{1-\alpha^{2h}}{1-\alpha}
\end{aligned}$$

where

$$\begin{aligned}
A(h,\lambda,\alpha) &= \lambda\frac{-h\alpha^{h-1} + h\alpha^h + 1 - \alpha^h}{(1-\alpha)^2}, \\
E(X_T) &= \mu = \frac{\lambda}{1-\alpha}, E(X_T^2) = \frac{\lambda(\lambda + 1 - \alpha)}{(1-\alpha)^2}.
\end{aligned}$$

4 Bootstrap forecast intervals for INAR(p) processes

Our proposal to obtain bootstrap prediction interval for INAR(p) process is as follows.

step 1 Compute the residuals: $\hat{\epsilon}_t = X_t - (\hat{\alpha}_1 X_{t-1} + \cdots + \hat{\alpha}_p X_{t-p}), t = p+1, \cdots, n.$, where $\hat{\alpha}_1, \cdots, \hat{\alpha}_p$ are the least squares estimator. Notice that the right hand side of the above equation is the difference between X_t and estimate of its conditional expectation(up to missing term $\hat{\mu}_\epsilon$).

step 2 Since each error $\hat{\epsilon}_t$ may be include a fractional part and even have a negative value, the considered empirical distribution is that of the modified errors $\tilde{\epsilon}_t$ defined by

$$\tilde{\epsilon}_t = \begin{cases} [\hat{\epsilon}_t] & : \hat{\epsilon}_t > 0 \\ 0 & : \hat{\epsilon}_t \le 0 \end{cases}$$

where $[\cdot]$ represents the value rounded to the nearest integer. Define the empirical distribution function of the modified residuals:

$$\hat{F}_{\tilde{\epsilon}}(x) = (n-p)^{-1} \sum_{t=p+1}^{n} 1_{\{\tilde{\epsilon}_t \le x\}}$$

step 3 Draw a i.i.d sample ϵ_t^* from the empirical distribution $\hat{F}_{\tilde{\epsilon}}(\cdot)$

step 4 Define X_t^* by the recursion :

$$\text{for, } t = 1, 2, \ldots, n, X_t^* = \sum_{i=1}^{p} \hat{\alpha}_i \circ X_{t-i}^* + \epsilon_t^*$$

step 5 Based on $\{X_1^*, X_2^*, \cdots, X_T^*\}, T \leq n$, compute the the estimation of the INAR(p) coefficients $(\hat{\alpha}_1^*, \ldots, \hat{\alpha}_p^*)$, as in step 2.

step 6 Compute future bootstrap observations by the recursion:

$$X_{T+h}^* = \sum_{i=1}^{p} \hat{\alpha}_i^* \circ X_{T+h-i}^* + \epsilon_t^*.$$

where $h > 0$, and $X_t^* = X_t$, for $t \leq T$.

The differences between our method and Alonso et al.(2002)'s method are step 2. In step 2, the residual $\hat{\epsilon}_t$ is neither centered nor rescaled. Since the INAR(p) model includes the disturbance term ϵ_t with mean μ_ϵ which is not zero, a centering procedure for $\hat{\epsilon}_t$ is not needed in step 2. Also, residual $\hat{\epsilon}_t$ may include a fractional part, so we considered modified residual $\tilde{\epsilon}_t$. When we omit step 5, the algorithm above is the conditional sieve bootstrap(CS) of Cao et al.(1997) and, otherwise, it is the complete sieve bootstrap(VS). The main difference of CS from VS is step 5 which incorporate the variability due to parameter estimation.

5 Simulation results

In this section the results of a simulation study of the bootstrap prediction intervals are presented. The true content of the intervals for finite samples is affected by parameter estimation method, the nature of error distribution, and the form of the model. We limit attention to the following model.

$< model1. > X_t = \alpha_1 \circ X_{t-1} + \epsilon_t \quad < model2. > X_t = \alpha_1 \circ X_{t-1} + \alpha_2 \circ X_{t-2} + \epsilon_t$

The error distributions F_ϵ considered are the Poisson, Negative binomial distribution which has small mean 3, 5, 10. Especially, when $\epsilon_t \sim Poisson(\lambda)$, the marginal distribution of the INAR(1) process is $Poisson(\lambda/(1-\alpha))$. The Poisson error represents the equidispersed case, and Negative binomial error represents overdispersed case.

Since the condition for stationarity for INAR(p) process is the same for AR(p)(Du and Li, 1991), we take $\alpha_1 = 0.3, 0.5, 0.7$ in model 1 and suitable combinations α_1, α_2 in model 2.

We take sample size $n = 25$ and 50 , leads $h = 1, 2, 3, 4, 5$, and nominal coverage $1 - \alpha = 0.95$.

To compare the different prediction intervals, we use their mean coverage and length, the proportions of observations lying out to the left and to the right of the interval and combined measure of coverage and length. We do the followings.

1. Simulate a series of specified structure, length, and error distribution, and generate $R = 1000$ true future values X_{T+k} from that series, using F_ϵ, the true parameter values.

2. For each bootstrap procedure obtain the $(1-\alpha)\%$ prediction interval $[Q_M^*(\alpha/2), Q_M^*(1-\alpha/2)]$ based on $B = 1000$ bootstrap resamples.

3. The coverage for each method is estimated as

$$C_M = \frac{\sharp\{Q_M^*(\alpha/2) \leq X_{T+h}^r \leq Q_M^*(1-\alpha/2)\}}{R},$$

where $X_{T+h}^r (r = 1, 2, \cdots, R)$ are the future values generated in the first step.

In step 1, we get "empirical" interval lengths using $L_T = X_{T+h}^{[R(1-\alpha/2)]} - X_{T+h}^{[R\alpha/2)]}$ and in step 2, bootstrap interval lengths using $L_M = [Q_M^*(\alpha/2), Q_M^*(1-\alpha/2)]$.

Steps 1-3 are then repeated $S = 200$ times to get $C_{M,i}, L_{M,i}$, $i = 1, \cdots, S$, and we get summary measures:

$$\begin{aligned} \bar{C}_M &= \sum_{i=1}^{S} C_{M,i}/S \qquad (5) \\ SE(\bar{C}_M) &= \left[\sum_{i=1}^{S}(C_{M,i} - \bar{C}_M)^2/(S(S-1))\right]^{1/2} \\ \bar{L}_M &= \sum_{i=1}^{S} L_{M,i}/S \\ SE(\bar{L}_M) &= \left[\sum_{i=1}^{S}(L_{M,i} - \bar{L}_M)^2/(S(S-1))\right]^{1/2} \\ CQ_M &= |1 - \bar{C}_M/\bar{C}_T| + |1 - \bar{L}_M/\bar{L}_T| \end{aligned}$$

where $\bar{L}_T = \sum_{i=1}^{S} L_{T,i}/S$ is the estimated true mean interval length, $\bar{C}_T = (1-\alpha)\%$ is the nominal coverage.

Table 1: Monte Carlo results for model 1, with poisson errors mean 3 and $\alpha_1 = 0.3$

Lead	Sample size	Method	$C_M(se)$	Cov (below/above)	$L_M(se)$	CQ_M
h	n	Theoretical	0.95	0.025/0.025	7.58	0
1	25	CS	0.95752(0.0030075)	0.004375/0.038105	7.775(0.0967921)	0.0336414
		VS	0.96687(0.0022903)	0.000315/0.032815	8.185(0.108224)	0.0975732
	50	CS	0.97159 (0.001701)	0.003175/0.025235	7.9(0.0690932)	0.0704717
		VS	0.974155 (0.0014179)	0.000865/0.02498	8.145(0.0800118)	0.105665
h	n	Theoretical	0.95	0.025/0.025	7.88	0
3	25	CS	0.94814 (0.0035168)	0.001005 /0.050855	7.915(0.1014412)	0.0063995
		VS	0.96519(0.0025576)	0.00065/0.03416	8.425 (0.0952219)	0.0851519
	50	CS	0.968155 (0.0018225)	0.00087 /0.030975	8.26 (0.0677121)	0.0720105
		VS	0.97258 (0.0015028)	0.00038 / 0.02704	8.445 (0.0678816)	0.1002503
h	n	Theoretical	0.95	0.025/0.025	7.88	0
5	25	CS	0.9492 (0.0037737)	0.00074 /0.05006	7.955 (0.1000747)	0.0103599
		VS	0.968515 (0.0023938)	0.00037 /0.031115	8.545 (0.090725)	0.1038803
	50	CS	0.966385 (0.0019244)	0.00079 /0.032825	8.205 (0.0681033)	0.0591521
		VS	0.97273 (0.0016255)	0.000465 /0.026805	8.49 (0.0657351)	0.1020216

Standard errors(SE) are in parentheses.

The below/above values give the ratio of the future values that fall below or above the interval ($\bar{C}_M$ =1 -below -above)

References

Al-Osh, M. A., and Alzaid, A. A. (1987), "First-order integer-valued autoregressive (INAR(1)) process," *Journal of Time Series Analysis*, Vol. 8, No. 3, 261-275.

Alzaid, A. A., and Al-Osh, M. (1990), "An integer-valued *p*th-order autoregressive structure (INAR(*p*)) process," *Journal of Applied Probability*, 27, 314-324.

Du, J.-Guan and Li, Y. (1991), "The integer-valued autoregressive (INAR(*p*)) model," *Journal of Time Series Analysis*, 12, 129-142.

Sueutel, F. W., and Van Harn, K. (1979), "Discrete analogues of self-decomposability and stability," *Annals of Probability*, 7, 893-899.

Stine, R. A. (1987), "Estimating properties of autoregressive forecasts," *Journal of American Statistical Association*, 82, 1072-1078.

Thombs, L. A., and Schucany, W. R. (1990), "Bootstrap prediction intervals for autoregression," *Journal of American Statistical Association*, 85, 486-492.

Alonso, A. M., Peña, D., and Romo, J. (2002), "Forecasting time series with sieve bootstrap," *Journal of Statistical Planning and Inference*, 100, 1-11.

Bloomfield, P. (1972), "On the error of prediction of a time series," *Biometrika*, 59, 501-507.

Bhansali, R. J. (1974), "Asymptotic mean-square error of predicting more than one-step ahead using the regression method," *Journal of Applied Statistics*, 23, 35-42.

Schmidt, P. (1974), "The asymptotic distribution of forecasts in the dynamic simulation of economic model," *Econometrika*, 42, 303-309.

Yamamoto, T. (1976), "Asymptotic mean square prediction error for an autoregressive model with estimated coefficients," *Journal of Applied Statistics*, 25, 123-127.

Freeland, R. K., and McCabe, B. P. M. (2004),"Forecasting discrete valued low count time series," *International Journal of Forecasting*, 20, 427-434.