Available online at www.sciencedirect.com
ScienceDirect
International Journal of Forecasting 24 (2008) 151 – 162
www.elsevier.com/locate/ijforecast

Model selection, estimation and forecasting in INAR(p) models: A likelihood-based Markov Chain approach

Ruijun Bu, Brendan McCabe *

The University of Liverpool, UK

Abstract

This paper considers model selection, estimation and forecasting for a class of integer autoregressive models suitable for use when analysing time series count data. Any number of lags may be entertained, and estimation may be performed by likelihood methods. Model selection is enhanced by the use of new residual processes that are defined for each of the $p+1$ unobserved components of the model. Forecasts are produced by treating the model as a Markov Chain, and estimation error is accounted for by providing confidence intervals for the probabilities of each member of the support of the count data variable. Confidence intervals are also available for more complicated event forecasts such as functions of the cumulative distribution function, e.g., for probabilities that the future count will exceed a given threshold. A data set of Australian counts on medical injuries is analysed in detail.

Keywords: Time Series of Counts; INAR(p) models; Maximum Likelihood Estimation; Markov Chain; Transition Probability; Transition Matrix; Delta Method

1. Introduction

One of the objectives of modelling time series data is to forecast future values of the variables of interest. The most common procedure for constructing forecasts in time series models is to use conditional expectations, as this technique will yield forecasts with the minimum mean squared forecast error. However, this method will invariably produce non-integer-valued forecasts, which are thus deemed to lack data coherency in the context of count data models. This paper presents a method of coherent forecasting for count data time series based on the integer autoregressive, or INAR(p), class of models. Integer autoregressive models were introduced by Al-Osh and Alzaid (1987) and McKenzie (1988) for models with one lag. Both Alzaid and Al-Osh (1990) and Du and Li (1991) also considered the INAR(p) class, but with different specifications of the thinning operators. In this paper we use the conditionally independent thinning scheme of Du and Li (1991). Freeland and McCabe (2004b) suggest using the h-step ahead conditional distribution and its median to generate data coherent predictions in the INAR(1) case. They also suggest that the probabilities associated with each point mass be modified to reflect the variation in

* Corresponding author. The School of Management, Chatham Street, Liverpool L69 7ZH, UK. Tel.: +44 151 795 3705; fax: +44 151 795 3005.

E-mail address: Brendan.McCabe@liv.ac.uk (B. McCabe).

0169-2070/$ - see front matter © 2007 Published by Elsevier B.V. on behalf of International Institute of Forecasters.
doi:10.1016/j.ijforecast.2007.11.002

parameter estimation. McCabe and Martin (2005) explored the issue of coherent forecasting with count data models under the Bayesian framework, but they too are concerned only with the first-order case. More recently, Jung and Tremayne (2006) proposed a simulation-based method for producing coherent forecasts for higher-order INAR models, but this too requires considerable computational work and does not use likelihood methods.

This paper makes three contributions. First, we suggest that the model be estimated by Maximum Likelihood (ML), should distributional assumptions warrant it.[1] We may therefore take advantage of the well known asymptotic normality and efficiency properties of the ML method. ML is not computationally difficult, and it allows a richer set of tools for model selection and improvement than do other methods of estimation for this class of models. For example, consider testing whether or not a thinning component should be excluded from the model; i.e., testing whether the associated parameter $\alpha_k=0$. Since α_k is a probability, methods of estimation require that $\hat{\alpha}_k$ be restricted to [0,1), and so tests based on $\hat{\alpha}_k$ will have a non-standard distribution because of the truncation at the boundary point 0. This truncation is not an issue for score-based tests in the ML framework. Other techniques like multiple residual analysis and specification testing are also available in the ML framework. Moreover, not only is the model estimated by ML, so too is the *entire h-step-ahead probability mass function*. This provides this method of forecasting with an optimality property. Estimation uncertainty can be accommodated by computing confidence intervals for these probabilities. Secondly, we suggest that the forecast mass function be computed by using a Markov Chain (MC) representation of the model. This method, while simple, avoids the need to evaluate complicated convolutions, and the same technique may be applied to any arrivals distribution and thinning mechanism. Thirdly, we consider forecasting the *cumulative distribution function* and events based on it. While it is undoubtedly interesting to know what the probability distribution of the size of a queue is, it is often more important to know what the probability is that the number will exceed a certain critical threshold. This requires forecasts of the cumulative distribution function and confidence intervals for the associated probabilities. This paper explains how confidence intervals with the correct coverage may be constructed.

A data set consisting of counts of deaths (by medical injury), monthly from January 1997 to December 2003, is analysed by ML techniques. Lag selection is achieved by means of residuals analysis and specification tests. The selected model is used to forecast up to 8 months ahead. Forecasts are made for both the probability mass and the cumulative distribution functions.

The remainder of the paper is organized as follows. Section 2 outlines the INAR(p) model and briefly discusses its properties. In Section 3, we present a method for producing h-step-ahead forecasts of the conditional probability distribution of the INAR(p) process. We also show how parameter uncertainty can be reflected in confidence intervals for probability forecasts. The medical injury death count data is analysed in Section 4, and Section 5 concludes.

2. The INAR(p) Model

Du and Li (1991) define the INAR(p) model to be

$$X_t = \alpha_1 \circ X_{t-1} + \alpha_2 \circ X_{t-2} + \cdots + \alpha_p \circ X_{t-p} + \varepsilon_t, \qquad (1)$$

where the innovation process $\{\varepsilon_t\}$ is i.i.d. (μ_ε, σ_ε^2) and is assumed to be independent of all thinning operations $\alpha_k \circ X_{t-k}$ for k=1,2,…, p, which are, in turn, conditionally independent. The "$\circ$" is the thinning operator, which, conditional on X_{t-k}, is defined as

$$\alpha_k \circ X_{t-k} = \sum_{i=1}^{X_{t-k}} B_{i,k},$$

where each collection $\{B_{i,k},\ i=1,2,\ldots,\ X_{t-k}\}$ consists of independently distributed Bernoulli random variables with parameter α_k, and the collections are mutually independent for $k=1,2,\ldots,\ p$. Intuitively, $\alpha_k \circ X_{t-k}$ is the number of individuals that would independently survive a binomial experiment in a given period, where each of the X_{t-k} individuals has an identical surviving probability α_k. The case where $p=1$ and $\{\varepsilon_t\}$ is known as Poisson autoregression, often denoted by

[1] Of course, the INAR model with Poisson arrivals could be used as a pseudo-likelihood, with the appropriate "sandwich" modification to the usual standard errors. We do not follow up on this suggestion here.