MODELING TIME SERIES OF COUNTS UNDER CENSORING AND TRUNCATION*

ISABEL PEREIRA, MARIA EDUARDA SILVA

ABSTRACT: *Censored and truncated data are frequently encountered in diverse fields including environmental monitoring, medicine, economics and social sciences. Censoring occurs when observations are available only for a restricted range, e.g., due to a detection limit. Truncation, on the other hand, occurs if observations in some range are lost. This work considers the analysis of time series of counts under censoring and truncation based on first order integer autoregressive (INAR) models. The focus is on estimation and inference problems.*

KEYWORDS: *Censored count series, INAR model, Truncated time series*

1 Introduction

Observations collected over time or space are often autocorrelated rather than independent. Time series measurements are often observed with data irregularities, e.g., observations with a detection limit. For instance, a monitoring device usually has a detection limit and it records the limit value when the true value exceeds/precedes the detection limit. This is often called censoring. So, censoring occurs if observations Y_i are available for a restricted range of Y_i arising from aggregation or may be imposed by survey design. This is very common in various situations in physical sciences, business, and economics. However the dependent variables can be limited in their range as well by truncation. Truncation arises if observations Y_i in some range of Y_i are totally lost.However the dependent variables can be limited in their range by truncation. Truncation arises if observations in some range are totally lost. Examples of left truncation include the number of bus trips made per week in surveys taken on buses, the number of shopping trips made by individuals sampled at a mall, and the number of unemployment spells among a pool of unemployed. Right truncation occurs if high counts are not observed. The most common form of truncation in count models is left truncation at zero. There is an extensive literature on regression models in which dependent variables are censored or underlying distributions are truncated, see [3] for a comprehensive review including discrete dependent variables. Nevertheless, the analysis of time series under censoring or truncation has received little attention in the literature and regard continuous valued processes only, see [4] and [1].

The model considered in this work is a time series of counts $X_1,\ldots,X_T$, generated according to the first order integer autoregressive process

$$X_t = \alpha \diamond X_{t-1} + \varepsilon_t \tag{1}$$

where the arrival process ε_t is a sequence of independent and identically distributed non-negative integer valued random variables, independent of X_{t-1}, with mean μ_ε and finite variance σ^2_ε and, conditional on X_{t-1}, $\alpha \diamond X_{t-1}$ is an integer valued random variable whose probability distribution, denoted by $g(\cdot|x;\alpha)$ depends on the parameter α which may be a vector of parameters. Thus, $'\diamond'$ denotes a random operator, usually called thinning operator, which always produces integer values and introduces serial dependence via the conditioning on X_{t-1}. For a review of thinning

*This work was supported in part by the Portuguese Foundation for Science and Technology (FCT-Fundação para a Ciência e a Tecnologia), through CIDMA - Center for Research and Development in Mathematics and Applications, within project UID/MAT/04106/2013.

operators see [2]. The specifications of the thinning operator and arrival process lead to INAR(1) models that ensure required distributional properties of the marginal distributions, namely that the marginal distribution of X_t is from the same family as ε_t. The transition probabilities are given by

$$p(X_t|X_{t-1}) = P[X_t = k|X_{t-1} = l] = \sum_{j=0}^{\min\{k,l\}} g(j|l)P[\varepsilon_t = k-j]. \tag{2}$$

Considering time series of counts based on first order integer autoregressive models this work aims at giving a contribution towards this direction.

2 A truncated count model

For some reason we do not observe X_t but Y_t which results from left truncating X_t at a value N. This means that a subset of the population that generated the data is unobserved. Then we define a **left truncated at N model**, for Y_t as

$$\begin{aligned} X_t &= \alpha \diamond X_{t-1} + \varepsilon_t \\ Y_t &= X_t, \text{if } X_t \geq N \end{aligned} \tag{3}$$

The transition probabilities are now

$$P[Y_t = y_t|Y_{t-1} = y_{t-1}] = \frac{P[X_t = y_t|X_{t-1} = y_{t-1}]}{\sum_{i=N}^{\infty}\sum_{j=N}^{\infty} P[X_t = j|X_{t-1} = i]\, P^*[X_{t-1} = i]} \tag{4}$$

with $P^*[X_t = j] = \frac{P[X_t=j]}{\sum_{i=N}^{\infty} P[X_t=i]}$ and zero outside $y_t, y_{t-1} \in [N, \infty[$.

3 A censored count model

Censoring occurs if observations Y_t are available for a restricted range due to aggregation or detection limits. We can say that censored data are âĂŸâĂŸpiled upâĂŹâĂŹ at a censoring point. We define a **censored at N model** as

$$\begin{aligned} X_t &= \alpha \diamond (X_{t-1}) + \varepsilon_t \\ Y_t &= \min\{X_t, N\} = \begin{cases} X_t, & \text{if } X_t \leq N \\ N, & \text{if } X_t > N \end{cases} \end{aligned} \tag{5}$$

The transition probabilities are

$$
\begin{aligned}
&P\left[Y_t = y_t | Y_{t-1} = y_{t-1}\right] = \\
&\begin{cases}
P\left[X_t \geq N | X_{t-1} \geq N\right] = \\
\sum_{i=N}^{+\infty} \sum_{j=N}^{+\infty} P\left[X_t = j | X_{t-1} = i\right] P^*\left[X_{t-1} = i\right], & y_t, y_{t-1} = N \\
\\
P\left[X_t = y_t | X_{t-1} \geq N\right] = \\
\sum_{i=N}^{+\infty} P\left[X_t = y_t | X_{t-1} = i\right] P^*\left[X_{t-1} = i\right], & y_t < N, y_{t-1} = N \\
\\
P\left[X_t \geq N | X_{t-1} = y_{t-1}\right] = \\
\sum_{j=N}^{+\infty} P\left[X_t = j | X_{t-1} = y_{t-1}\right], & y_t = N, y_{t-1} < N \\
\\
\frac{P[X_t = y_t | X_{t-1} = y_{t-1}]}{P[X_t < N | X_{t-1} < N]} = \\
\frac{P[X_t = y_t | X_{t-1} = y_{t-1}]}{\sum_{i=N}^{+\infty} \sum_{j=0}^{N} P[X_t = j | X_{t-1} = i] P^*[X_{t-1} = i]}, & y_t, y_{t-1} < N
\end{cases}
\end{aligned}
\tag{6}
$$

4 Estimation procedures

Neglecting censoring and truncation in the time series hinders meaningful statistical inference, leading to model misspecification, biased parameter estimation, and poor forecasts. We study the regression properties of the censored INAR(1) models and show that least squares estimation of the parameters is no longer appropriate. Likelihood analysis of the censored INAR(1) processes is developed, including maximum likelihood and maximum pseudo-likelihood estimations. Some problems related to the censored or truncated count data models are pointed out and illustrated.

REFERENCES:

[1] Choi, Seokwoo Jake and Portnoy, Stephen (2016). Quantile Autoregression for Censored Data. *J. Time. Ser. Anal.*,37, pp. 603âĂŞ623.

[2] Gouveia, Sónia and Scotto, Manuel and Weiss, Christian H. and Ferreira, Paulo J.S.G. (2016). Binary autoregressive geometric modelling in a DNA context. *Journal of the Royal Statistical Society, Series C*.

[3] Greene, W. (2005). Censored Data and Truncated Distributions, *Working Papers 05-08*, New York University, Leonard N. Stern School of Business, Department of Economics.

[4] Park, Jung Wook and Genton, Marc G. and Ghosh, Sujit K. (2007). Censored time series analysis with autoregressive moving average models. *The Canadian Journal of Statistics*, 35, pp. 151-168.

Isabel Pereira
University of Aveiro & CIDMA, Portugal
E-mail:isabel.pereira@ua.pt

Maria Eduarda Silva
University of Porto & CIDMA, Portugal
E-mail:me.pintosilva@gmail.com