Proceedings of the 8th WSEAS International Conference on APPLIED MATHEMATICS, Tenerife, Spain, December 16-18, 2005 (pp363-366)

On AR(1) versus MA(1) Models for Non-stationary Time Series of Poisson Counts: Part II (Application)

VANDNA JOWAHEER
University of Mauritius
Department of Mathematics
Reduit, Mauritius
MAURITIUS

BRAJENDRA C. SUTRADHAR
Memorial University of Newfoundland
Department of Mathematics and Statistics
St. John's, NL, A1C5S7
CANADA

Abstract: Observations driven non-stationary Poisson AR(1) and MA(1) models may be applied to analyze various biomedical and socio-economic (e.g., monthly tourist counts) time series of counts. This paper fits such AR(1) and MA(1) models to the US monthly polio count data which was analyzed earlier by Zeger [7] and Davis, Dunsmuir and Wang [1] by using a random effects based correlated count response model. The goodness of fitting of the AR(1) and MA(1) models to this polio data set is also discussed.

Key–Words: Efficient generalized quasilikelihood (GQL) approach; MA(1) models for counts; Polio count data analysis.

1 Introduction

Analysis of time series of counts is an important research topic in the biomedical and socio-economic fields. The statistical analysis of this type of data has not been, however, easy, because of the difficulty in modelling the correlation structure of repeated counts over a long period of time. To model the correlation structure of the count data recorded over time, some authors such as Zeger [7] (see also Harvey and Fernandes [2], Davis et al [1]) assumes that conditional on a dependent sequence of a stationary Gaussian random effects with auto-correlation structure, the time series data follow independent Poisson distributions so that unconditionally the observations are correlated. Note however that in this approach, even though the random effects have a Gaussian correlation structure, they however yield a complicated correlation structure for the count responses.

Following the observations driven stationary AR(1) models (McKenzie [4,5]) for the count data, recently Mallick and Sutradhar [3] have extended the stationary AR(1) models for the negative binomial counts to the non-stationary case that include the non-stationary Poisson case as a sub-model. One may similarly construct an MA(1) non-stationary correlations model for the count data. This paper fits both of these AR(1) and MA(1) models to the US polio count data which was earlier analyzed by Zeger [7] and Davis et al [1]. To be specific, in Section 2, the AR(1) and MA(1) models are described in brief along with the GQL estimating equations for the parameters of these models. In Section 3, the estimation techniques from Section 2 has been applied to analyze the US polio data. Goodness of fit of these two models to the US polio data is also discussed.

2 Estimating Formulas Under Both AR(1) and MA(1) Models

Let y_t $(t = 1, \cdots, T)$ be the count response recorded at time t and x_t $(t = 1, \cdots, T)$ be the corresponding $p \times 1$ vector of covariates. Further let $\beta = (\beta_1, \cdots, \beta_p)'$ be the p-dimensional vector of regression effects. The GQL estimating equations for the regression effects and the moment estimating equations for the correlation parameters under both AR(1) and MA(1) models are given below.

2.1 Non-stationary AR(1) Models

Model: $y_t = \rho * y_{t-1} + d_t$,
where $y_{t-1} \sim P(\mu_{t-1}),\ d_t \sim P(\mu_t - \rho\mu_{t-1})$, with $\mu_t = e^{x_t'\beta}$, and $\rho * y_{t-1} = \sum_{j=1}^{y_{t-1}} b_j(\rho)$, with $b_j(\rho)$ representing a binary variable such that $pr(b_j(\rho) = 1) = \rho$ and $pr(b_j(\rho) = 0) = 1 - \rho$. Also y_{t-1} and d_t are independent.

GQL Estimating Equation for β: Let $y = (y_1, \cdots, y_t, \cdots, y_T)'$ be the T-dimensional vector of all responses and $\mu = (\mu_1, \cdots, \mu_t, \cdots, \mu_T)'$ be the mean vector of y, and $\Sigma = (\sigma_{tt'})$ be the $T \times T$ covariance matrix of y, where $\sigma_{tt'} =$

$\left(\sqrt{\{\text{var}(Y_t)\text{var}(Y_{t'})\}}\rho_y(\ell)\right)$, with $\text{var}(Y_t)$ and $\rho_y(\ell)$ representing the variance and lag ℓ correlations of the responses. Then the GQL estimate of β is obtained by solving the estimating equation $\frac{\partial\mu'}{\partial\beta}\Sigma^{-1}(y-\mu)=0$.

Moment Equation for ρ: Under the AR(1) model, the moment estimate of ρ has the formula given by

$$\hat{\rho} = \frac{\sum_{t=2}^{T}\tilde{y}_t\tilde{y}_{t-1}}{\sum_{t=1}^{T}\tilde{y}_t^2}\frac{T}{\sum_{t=2}^{T}[\mu_{t-1}/\mu_t]^{\frac{1}{2}}},$$

where $\tilde{y}_t = [y_t-\mu_t]/\sqrt{\mu_t}$.

2.2 Non-stationary MA(1) Model

Model: $y_t = \rho * d_{t-1} + d_t$,
where $d_t \sim P(\mu_t/(1+\rho))$, and $d_{t-1} \sim P(\mu_{t-1}/(1+\rho))$, respectively.

The GQL Estimating Equation for β: Let $\nu = (\nu_1,\cdots,\nu_t,\cdots,\nu_T)'$ be the mean vector of $y = (y_1,\cdots,y_t,\cdots,y_T)'$, with $\nu_t = E(Y_t) = \text{var}(Y_t)$ for all $t = 1,\ldots,T$. Furthermore, let $\tilde{\Sigma} = (\sigma_{tt'})$ be the $T\times T$ covariance matrix of y. Then for known ρ, the GQL estimating equation for β is given by $\frac{\partial\nu'}{\partial\beta}\tilde{\Sigma}^{-1}(y-\nu) = 0$, (Sutradhar [6]) which may be solved iteratively by Newton-Raphson iterative technique.

Moment Estimating Equation for the Correlation Parameter ρ: A consistent estimate of ρ is obtained by solving the moment equation derived from the covariance equation given by

$$E\left[(Y_t-\nu_t)(Y_{t-1}-\nu_{t-1})\right] = \{\rho/(1+\rho)\}\mu_{t-1}.$$

3 Analysis of U.S. Polio Count Data

This section first fits the non-stationary Poisson MA(1) model to the time series of the monthly number of cases of poliomyelitis reported by the U.S. Centers for Disease Control for the years 1970-1983. Here total number of observations is $T = 168$. For convenience the data set is shown in Figure 1. Note that this data was first analyzed by Zeger [7] and then by Davis et al [1] both by using their proposed random effects based parameter driven models. For the purpose of comparison, the same regression variables as in Zeger [7] have been used. Consequently, the monthly number of polio cases are regressed on a linear trend as well as sine, cosine pairs at annual and semi-annual frequencies to reveal the evidence of seasonality. More specifically, the selected covariates are :

$$x_t = \Big[1, t'/1000, cos(2\pi t'/12), sin(2\pi t'/12), cos(2\pi t'/6), cos(2\pi t'/6)]', \Big]$$

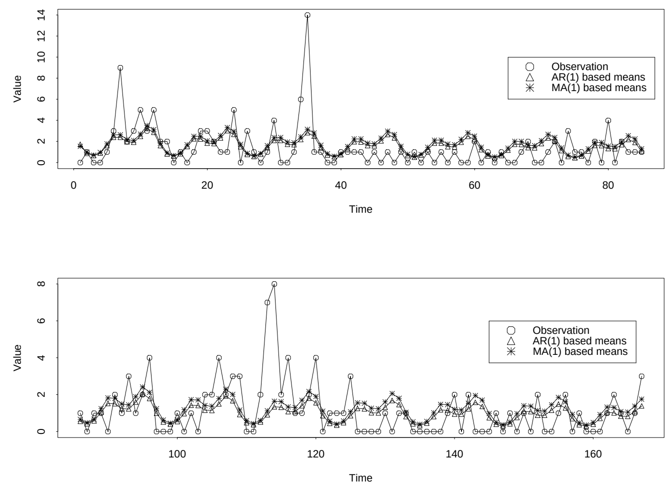

Figure 1: U.S. polio count data from January 1970 to December 1983 and expected counts based on AR(1) and MA(1) models.

where $t' = (t-73)$ is used to locate the intercept term at January 1976, for $t = 1,\cdots,168$. Mallick and Sutradhar (2005) have analyzed the same polio data by using the non-stationary AR(1) model for the count data. In this section, a comparison of the inferences is made by fitting the AR(1) and MA(1) models.

Under the non-stationary MA(1) model, the β and the ρ parameters are estimated simultaneously by using the formulas from Section 2.2. The standard errors of the regression estimates are computed by an appropriate formula derived from the estimating equation. The regression estimates along with their standard errors as well as the estimate of ρ parameter are given in Table 1.

Table 1. Comparison of the estimates of the regression and the correlation parameters under the AR(1) and MA(1) models in fitting the original as well as the modified (adjusted for possible outliers) polio data .

Fitting original data

	AR(1) model		MA(1) Model	
Parameters	EST	SE	EST	SE
Intercept (β_1)	0.19	0.094	0.33	0.469
Trend$\times 10^{-3}$ (β_2)	-5.49	1.792	-4.47	-
$\cos(2\pi t/12)(\beta_3)$	-0.19	0.122	-0.28	0.096
$\sin(2\pi t/12)(\beta_4)$	-0.52	0.127	-0.52	0.072
$\cos(2\pi t/6)(\beta_5)$	0.13	0.106	0.04	0.088
$\sin(2\pi t/6)(\beta_6)$	-0.41	0.110	-0.49	0.092
ρ	0.22	–	0.32	–
$\rho_y(1)$	0.23	–	0.24	–
SSD	309.5	–	268.5	–

Fitting modified data

	AR(1) model		MA(1) Model	
Parameters	EST	SE	EST	SE
Intercept (β_1)	0.03	0.084	0.02	0.122
Trend$\times 10^{-3}$ (β_2)	-4.12	1.582	-3.84	1.490
$\cos(2\pi t/12)(\beta_3)$	-0.03	0.105	-0.03	0.105
$\sin(2\pi t/12)(\beta_4)$	-0.50	0.120	-0.49	0.120
$\cos(2\pi t/6)(\beta_5)$	0.21	0.109	0.21	0.109
$\sin(2\pi t/6)(\beta_6)$	-0.18	0.109	-0.19	0.109
ρ	0.007	–	0.008	–
$\rho_y(1)$	0.007	–	0.008	–
SSD	177.0	–	178.1	–

The estimates of these parameters under the AR(1) model are computed based on the formulas given in Section 2.1. These results are also given in the same table. The ρ parameter under the MA(1) model is estimates as 0.32, whereas under the AR(1) model, this estimate was found to be 0.22. They along with the estimates of β however appear to produce the almost same lag 1 correlation 0.23 or 0.24. Note that the lag 1 correlations are non-stationary under both AR(1) and MA(1) models, which are given in Figure 2 for clarity. The averages of these non-stationary correlations are referred to as the lag 1 correlations. The regression estimates appear to be different in general under the two models. Except for β_1 and β_2, the standard errors of the other regression estimates are smaller under the MA(1) model as compared to those of the AR(1) model, indicating that MA(1) model produces efficient regression estimates. Next to see the over all fit, the traditional standardized squared distances (SSD) have been computed under both AR(1) and MA(1) models. Under the AR(1) model this is given by SSD $= \sum_{t=1}^{T}\left[(y_t-\mu_t)/\sqrt{\mu_t}\right]^2$, whereas under the MA(1) model this SSD is given as SSD $= \sum_{t=1}^{T}\left[(y_t-\nu_t)/\sqrt{\nu_t}\right]^2$. The SSD under the MA(1) model was found to be 268.48, whereas the AR(1) model produced the SSD as 309.5. Consequently, the polio data considered here appears to be better fitted by the MA(1) model. As far as the regression estimates are concerned, most of the estimates including the trend and seasonal coefficients, are found to be negative indicating that the polio cases were decreasing in general over the years.

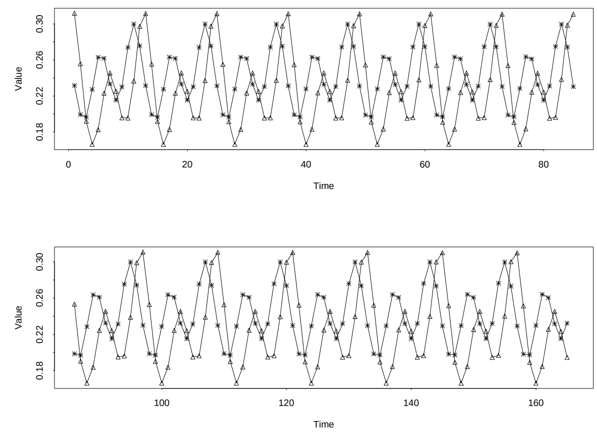

Figure 2: AR(1) and MA(1) models based non-stationary lag 1 correlations for the U.S. polio data from January 1970 to December 1983; $\triangle$ *stands for AR(1) based correlations and* $*$ *stands for MA(1) based correlations*.

Note that even though the AR(1) model fitted the data somewhat worse than the MA(1) model, there were no problems to obtain the the standard errors of the GQL estimates of the regression parameters under the AR(1) model, whereas the GQL estimate of the β_2 parameter produced negative variance estimate under the MA(1) model. This happened as the polio data appears to have a few larger counts (see Figure 1) such as $y_t = 9,\ 14,\ 7,\ 8$ at time points $t = 7,\ 35,\ 113,\ 114$ respectively. Thus, the GQL estimation approach under the MA(1) model unlike the AR(1) model may not be suitable in general to deal with outlier type observations. This became evident from a re-analysis by replacing these moderately large counts by the mean 1 of the rest of the data. The GQL estimates of the regression and correlation parameters for this modified data set are also given in Table 1 under both AR(1) and MA(1) models. It is clear from the table that both of these models produce the same estimates for the regression and the correlation parameters in fitting the modified data. Furthermore, the modified polio data appear to be almost independent over the years as the estimates of the correlation parameters were found to be 0.01. The goodness of fit by the two models were also to be the same. Thus, one may fit either of the two models to the modified polio data , where modification was done by down weighting the outlier type count observations.

4 Conclusion

This paper has demonstrated how one can fit the observations driven AR(1) and MA(1) models to a time series of non-stationary counts. This was done by fitting these models to a time series of polio counts from the USA. The present analysis also indicates that one would require a suitable robust estimation technique to fit any time series of counts in the presence of possible outliers. This generalization is however beyond the present article.

Acknowledgements: This research was partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The authors would like to thank the reviewers for helpful comments and suggestions.

References:

[1] R. A. Davis, W. T. M. Dunsmuir, and Y. Wang, On autocorrelation in a Poisson regression model, *Biometrika* 87, 2000, pp. 491-505.

[2] A. C. Harvey and C. Fernandes, Time series models for count or qualitative observations, *Journal of Business and Economic Statistics* 7, 1989, pp. 407-417.

[3] T. Mallick and B. C. Sutradhar, Analyzing time series of non-stationary negative binomial series, *Technical Report No. 4*, *Department of Mathematics and Statistics*, *Memorial University of Newfoundland*, *Canada*, 2005.

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

[5] E. McKenzie, Some ARMA models for dependent sequences of Poisson counts, *Adv. in Appl. Probab.* 20, 1988, pp. 822-835.

[6] B. C. Sutradhar, An overview on regression models for discrete longitudinal responses, *Statistical Science* 18, 2003, pp. 377-393.

[7] S. L. Zeger, A regression model for time series of counts, *Biometrika* 75, 1988, pp. 621-629.