Discrete valued time series models for examining weather effects in daily accident counts

Dimitris Karlis[1], George J. Sermaidis[2] and Tom Brijs[3]

[1] Department of Statistics, Athens University of Economics and Business, 76 Patission str., 10434 Athens, Greece, e-mail: karlis@aueb.gr
[2] Department of Statistics, University of Warwick, Coventry CV4 7AL, England
[3] Transportation Research Institute, Hasselt University, Wetenschapspark, Gebouw 5, B-3590 Diepenbeek, Belgium

Abstract: In this paper we aim at examining the effect of weather conditions in daily accident counts. In order to account for the serial correlation and the overdispersion present to the data, we make use of two models for discrete valued time series using covariate information. The models considered are the model of Zeger and the Integer Autoregressive model including covariates. Estimation procedures and possible extensions of the models are discussed. Data from 27 major cities roads in the Netherlands are examined. We make use of a meta-analysis approach in order to combine the effects retrieved for each site with site-specific covariate informtaion.

Keywords: INAR model, Zeger's model, accidents statistics

1 Introduction

The last few years, road accidents statistics are the subject of increased interest both on the part of policy makers and academia. The objective is to better understand the complexity of factors that are related to road accidents in order to take corrective actions to remedy this situation. In this context, the modelling of accidents over time has obtained considerable attention by researchers in the past. In this paper, we study the effects of weather conditions on daily accidents for 27 major cities in the Netherlands. The use of weather conditions is motivated by earlier research where significant influences of weather conditions on accidents have been found. To do so we propose and apply two models adequate for modelling discrete valued time series models, namely the Integer Autoregressive model proposed by McKenzie (1985) and Al-Osh and Al-Zaid (1987) and the model proposed by Zeger (1988). The first belongs to the category of observation driven models, since we relate directly the observations themselves, while the second one to the category of the parameter driven models as the de-

pendence structure comes from a time dependent process in the parameters of the model.
In order to capture the effect of weather covariates in the accident counts we use covariate information related to weather conditions. This includes covariate information about the rainfall, the wind, the temperature and other weather characteristics in the area of examination.
Furthermore in order to account for the different characteristics of the 27 roads we proceed with meta-analysis of the derived results. This approach allows to combine the results from the different sites and to examine site-specific effects.

2 The models

2.1 INAR model

McKenzie (1985) and Al-Osh and Al-Zaid (1987) defined a process for discrete data which mimics the standard autoregressive model for continuous data, called the Integer-valued autoregressive (INAR) process as follows:
A sequence of random variables $\{Y_t\}$ is an INAR(1) process if it satisfies a difference equation of the form

$$Y_t = \alpha \circ Y_{t-1} + R_t, \quad t = 1, 2, \ldots, \tag{1}$$

where R_t is the innovation term, which is a discrete random variable. According to the choice of the distribution of the innovations certain marginal properties can be deduced for the process. The operator " $\circ$ " denotes the binomial thinning operator defined by $\alpha \circ Y = \sum_{t=1}^{Y} Z_t$, where Z_t are independent Bernoulli random variables with $P(Z_t = 1) = \alpha = 1 - P(Z_t = 0)$, $\alpha \in [0, 1]$. Thus, conditional on Y_t, $\alpha \circ Y_t$ is a binomial random variable where Y_t denotes the number of trials and α the probability of success in each trial.
The basic ingredient of the INAR model is that it assumes that the realization of the process at time t is composed by two parts, the first one clearly relates to the previous observation, while the second one is independent from it and depends only on the current time point. Thus, the first part represents the influence of previous time periods while the innovation term captures the effects of the present time point. Although it is possible to incorporate higher-order lags into the model, we do not pursue them since their interpretation is not straightforward. More detail on such models can be found in Jung and Tremayne (2006).
Assuming that R_t follows a Poisson distribution the Poisson INAR model arises, which assumes that the marginal distribution of Y_t is a Poisson distribution. The simple Poisson INAR model can be extended to a Poisson

INAR regression model by adding covariates to both the innovation term and/or the autocorrelation parameter. The model then takes the form

$$\begin{aligned} Y_t &= \alpha_t \circ Y_{t-1} + R_t \\ R_t &\sim Poisson(\lambda_t) \\ \log \lambda_t &= \mathbf{z_t}^{'}\beta \\ \log\left(\frac{\alpha_t}{1-\alpha_t}\right) &= \mathbf{w_t}^{'}\gamma \end{aligned}$$

where $\mathbf{z}_t$ and $\mathbf{w}_t$ are vectors of covariates at time t for the innovation term and the autocorrelation parameter respectively while β and γ are the vector of the associated regression coefficients. Note that the covariate information for the two parts of the model are not necessarily the same. We have developed an EM type algorithm for fitting this model to real data making use of the convolution representation of the process. Details of the algorithm are omitted.
Extensions of the model to allow for overdispersion can be made by assuming an overdispersed innovation distribution.

2.2 Zeger's Model

We describe the model proposed by Zeger (1988). Let's suppose we have observed a time series of counts y_t, $t = 1, 2, ...T$, as well as a vector of covariates $\mathbf{x}_t$. Our goal is to describe $\mu_t = E(Y_t)$ as a function of the $p\times 1$ vector of covariates. Furthermore, assuming that the distribution of y_t is Poisson, that is $y_t \sim Poisson(\mu_t)$, where $\mu_t = exp\ (\mathbf{x}_t^{'}\mathbf{b})$, maximum likelihood method can be used to estimate the unknown vector of coefficients $\mathbf{b}$. In practice, quite often the sample variance exceeds the sample mean, providing evidence that an overdispersed relative to the Poisson distribution must be used. In this case quasi-likelihood methods which allow a variety of variance-mean relation is more appropriate.
Extensions of log-linear models which account for dependence are necessary to obtain valid inference about the relationship of y_t and $\mathbf{x}_t$. Zeger suggested that if ϵ_t is an unobservable noise process then the conditional distribution of y_t on ϵ_t is Poisson with mean equal to the product of the latent process value and the predictor as in a simple log-linear model. Therefore

$$Y_t \mid \epsilon_t \sim Poisson(\epsilon_t exp\ (\mathbf{x}_t^{'}\mathbf{b})) \tag{2}$$

Assume that ϵ_t is a non-negative time series with mean 1, autocovariance function $\gamma_\epsilon(h)$ and variance σ_ϵ^2. Letting $\delta_t = \log \epsilon_t$, then the conditional mean of Y_t on ϵ_t can be written as

$$u_t = exp\ (\mathbf{x}_t^{'}\mathbf{b} + \delta_t) \tag{3}$$

We assume $E(\exp(\delta_t)) = 1$. Unless the δ_t is a stationary Gaussian process, there is not an explicit relationship between the autocovariance functions of ϵ_t and δ_t.

For this model, the marginal variance of Y_t is greater than its marginal mean providing this way a degree of overdispersion which depends on the variance of the latent process σ_ϵ^2. Another interesting property of this model is that the form of the autocorrelation of the observed counts inherits its structure from that of the latent process. It is also true that even if there is no significant autocorrelation in y_t, it does not necessarily mean that autocorrelation is not present in ϵ_t either. This implies that the autocorrelation function of the observed count process will tend to underestimate that of the latent process, even in the simplest case where no regressors are present. Therefore, the latent process introduces both autocorrelation and overdispersion in Y_t. The interpretation of any element of the vector of coefficients $\mathbf{b}$ in the above model is the same as in a simple Poisson regression model.

Estimation of this model is not easy. The full likelihood of the model cannot be written easily as it is defined recursively. A GEE approach has been proposed by Zeger (1988) in order to estimate the parameters of the model. We have followed this approach.

Concluding this section, the two model, despite their different generation mechanism implied, have some more differences in the sense that the model of Zeger allows for overdispersion. In the sequel we applied both model to our data.

3 Meta-Analysis

Meta-analysis can be defined as the quantitative review and synthesis of the results of related but independent studies. By combing information over different studies, an integrated analysis will have more statistical power to detect a specific effect than an analysis based on only one study. When several studies have conflicting conclusions, a meta-analysis can be used to estimate an average effect. For an excellent review on meta-analysis the reader can refer to Normand (1999). In this paper we aim at combining results from different sites in order to synthesize a general effect.

A fixed-effects model assumes that each study summary statistic Y_i (in our case a regression coefficient summarizing the effect of a weather variable) is a realization from a population of study estimates with common mean θ. Let α be the central parameter of interest and assume there are $i = 1, 2, ..., k$ independent studies. Assume that Y_i is such that $E(Y_i) = \theta$ and let $Var(Y_i) = s_i^2$ be the variance of the summary statistic in the ith study. For moderately large study sizes, each Y_i should be normally distributed (by the central limit theorem) and approximately unbiased. Thus

$$Y_i \sim N(\alpha, s_i^2) \quad \text{for} \quad i = 1, 2, ..., k \tag{4}$$

and s_i^2 assumed known. The central parameter of interest is α which quantifies the average effect.
The random-effects model assumes that each study summary statistic Y_i is drawn from a distribution with a study-specific mean, α_i, and variance s_i^2.

$$Y_i \mid \alpha_i, s_i^2 \sim N(\alpha_i, s_i^2) \quad \text{for} \quad i = 1, 2, ..., k \tag{5}$$

Furthermore, each study-specific mean α_i is assumed to have been drawn from some superpopulation of effects with mean α and variance τ^2 with

$$\alpha_i \mid \alpha, \tau^2 \sim N(\alpha, \tau^2) \tag{6}$$

The parameters α and τ^2 are to referred as hyperparameters and represent, respectively, the average effect and inter-study variation. Thus we introduce one more level of variability.

4 Application

This study is based on the daily accident counts that were obtained from the major roads covered by the surface of 27 big cities in the Netherlands in the year 2001. The cities were selected based on two criteria: a) their proximity to some national weather stations in order to obtain accurate daily weather conditions and b)they were far enough apart in order to prevent that weather conditions would be identical for the different sites for too many of the observations.
For each site, the number of daily counts on accidents together with detailed weather information was collected. The day was used as a proxy of the different traffic volumes, i.e. as a proxy to the exposure. The data have several distinct features since for some roads the autocorrelation and the overdispersion varied considerably.
We make use of both the fixed and the random effects meta-analysis models. A typical forest plot can be seen in Figure 1 for the precipitation duration. One can see the different effects for each site and the combined estimator. This combined estimator shows that there is a positive effect of the precipitation duration.
The meta-regression models identified variables which influence the effect of the covariates. A summary of the main finding is that an increase of one unit of the maximum temperature decreases the effect of the mean temperature on the accidents and a decrease of a unit of the minimum temperature increases the temperature below zero effect. The effect of humidity covariate becomes stronger when combined with lower minimum temperatures. The rainfall intensity effect was related to the increase of the rainfall duration. However, one must interpret the findings with care since weather variables can also have some influence on the exposure. Hence, the effects found are not necessarily explicit on the accidents but they can be implicit through the increase/decrase of the exposure.

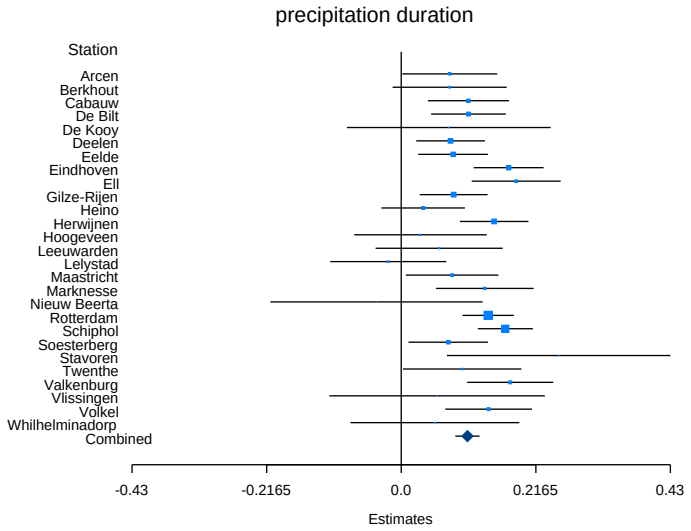

FIGURE 1. Weighted forest plot for the precipitation duration

References

Al-Osh M.A. and Al-Zaid A.A. (1987). First Order Integer Valued Autoregressive Process, *Journal of Time Series Analysis*,**8**, 261-275.

Jung, R.C. and Tremayne, A.R. (2006), Binomial thinning models for integer time series, *Statistical Modelling*, **6**, 81-96

McKenzie E. (1985). Some Simple Models for Discrete Variable Time Series. *Water Resources Bulletin*,**21**, 645-650.

Normand S.L. (1999) Meta-analysis: formulating, evaluating, combining, and reporting. *Statistics in Medicine***18**, 321-359.

Zeger S.L. (1988). A Regression Model for Time Series of Counts. *Biometrika*, **75(4)**, 621-9.