Modelling Time Series of Count Data

RICHARD A. DAVIS*
Colorado State University

WILLIAM T. M. DUNSMUIR
University of New South Wales

and

YING WANG
Colorado State University

ABSTRACT

The focus of this paper is on the similarities and differences in model building and interpretation for two types of conditional mean processes in Poisson regression models for time series of counts: parameter-driven and observation-driven specifications. For a parameter-driven model, it is shown that under general conditions, the Poisson maximum likelihood estimator of the regression parameter based on a model without serial correlation is consistent and asymptotically normal with an easily computable covariance matrix. This covariance matrix depends on the covariance structure of the latent process. A method of testing for the existence of a latent process is developed and compared to existing test statistics via simulation. Once the existence of a latent process has been detected, a simple and easily implementable method for estimating the autocorrelation of the latent process is given. The resulting estimates are shown to be consistent and the standard errors of the estimates are also provided. When the latent process is white noise, a test statistic for testing serial dependence is proposed based on the study of the distribution of autocorrelation estimates. Estimation of the regression and latent process parameters in a parameter-driven model is developed using an approximation to the likelihood. New and existing formulations of observation-driven models are also discussed. Properties of these observation-driven models are studied and likelihood based estimation procedures are presented. Issues surrounding model identification and estimation for both parameter and observation driven models are illustrated with two data sets.

*This research supported in part by NSF DMS Grant No. 9504596.
AMS 1980 Subject Classification: 62M10; 62M09; 62M07
Key Words and Phrases. GLM, Poisson regression, state-space models.

1 Introduction

1.1 Overview

In recent years there has been a growing literature on models and methods for analyzing time series of counts. The need for such a development is clearly demonstrated in applications such as modeling of disease incidence, as for example in the modeling of polio counts in U. S. (see Zeger, 1988). Another area that is increasingly important is the modeling of disease counts and their relationship to putative causal factors such as weather and atmospheric pollution. Numerous articles have appeared in this vein, for example Campbell (1994), Jorgensen et al. (1995), and papers by Schwartz et al. (1995, 1996). Another potential application of these methods is to time series of counts obtained from defect measurements on the output of some commercial or manufacturing process in which, as is the case for continuous output measurements, there is interest in relating process control variables to the outcome defect rate which could be a binomial count or a Poisson count. Other recent applications are to micro time scale financial data in which the rate at which transactions form during the day might be the variable of interest or to road traffic injuries or fatalities - see Brannas and Johansson (1994) or Harvey and Fernandes (1989) for example.

In the setting where the researcher wishes to relate disease outcomes to pollution, count measurements arise naturally because often the disease is relatively rare and because the time scale or geographical scale on which pollution measurements fluctuate are necessarily small. For example, in a study of the relationship between temperature and the incidence of sudden infant death syndrome (SIDS) Campbell worked with daily measurements of temperature and SIDS counts. If the disease is even rarer (e.g. polio) the time scale that is suitable might be one month. Aggregation over longer time periods in both of these examples would obliterate or at least blur the link between the putative risk factor and the disease outcome.

As in all time series modeling applications, the possibility of serial dependence within the data has to be investigated. In a linear model in which the response variable is continuous and possibly even Gaussian, the effects of serial correlation in the regression error structure have been well understood for about half a century. Likewise the various forms of model structure (serially correlated errors, lagged dependent variables forms, for example) are very well understood as is the theory of state-space models.

Only in the last 5 to 10 years has their been serious attention paid to the analogous results and models in the area of regression modeling for count data. The recent literature has seen various models and methods proposed and we will review some of these here.

However what is lacking, in our view, is a comprehensive comparison of these various models and methods in terms of their statistical properties and their performance on a variety of real data sets. Also, there is not yet a structured and statistically well founded approach to model building and diagnosis as is readily found in the literature for linear time series analysis. This article is aimed at making modest progress towards the goal of filling in these details or at the very least, identifying the need for additional theory and practice.

Analysis of time series of counts is concerned with the usual issues that arise in the standard linear modeling paradigm, namely: model fitting, hypothesis testing or relationship building, and forecasting. The use to which the model will be applied is important in the selection of model as we discuss below. In anticipation of that discussion we mention that models which are based on a presumption of a latent stochastic process governing the (conditional) mean function in the distribution of the observed count process are difficult to forecast with since the latent process is the means by which serial dependence is modeled and yet this is not directly observable. Furthermore, this is a non-Gaussian setting so that predictors are typically nonlinear. On the other hand models

which incorporate directly into their mean function lagged values of the dependent count process are easy to forecast with.

Reviews of the models considered in this paper can also be found in Brockwell and Davis (1996, Section 8.8) and Fahrmeir and Tutz (1994, Chapters 6-8). While the literature contains many suggestions, little comparative and analytical work has been carried out. Often asymptotic distributions, so necessary for valid inference (i.e. in which the effect of serial dependence on regression standard errors), are only loosely argued. For example, if a regression on time is performed (e.g. to build a linear trend term in the linear predictor) without normalization by the number of time periods the usual asymptotics do not apply when the trend is negative simply because the information matrix becomes degenerate.

A proper treatment of estimation of serial dependence has not yet been given. We provide some contributions to this below. Typically in Poisson regression the use of the residuals from the fit (based on observed counts minus the fitted values) will seriously underestimate the true serial dependence. We summarize existing and propose improved techniques for avoiding this underestimation. In addition we compute standard errors and analogue of the Box-Pierce portmanteau test for serial correlation widely used in linear time series models.

1.2 Desiderata of Models

Zeger and Qaqish (1988) implicitly offer desiderata that observation driven models should satisfy. For specificity, we assume that the count data, denoted by Y_t, follow a Poisson distribution with mean, conditional upon the r-dimensional regression variable x_t, given by

$$\mu_t = \exp(\mathbf{x}_t^T \boldsymbol{\beta} + \nu_t)$$

in which ν_t is a random process possibly dependent on lagged values of the observed counts Y_{t-l}. It is assumed that the joint distribution of the ν_t depends on a vector of parameters $\boldsymbol{\gamma}$. If $\nu_t \equiv 0$, then this is the familiar Poisson regression model. The three desiderata described in Zeger and Qaqish (1988) are:

D.1. The marginal mean of Y_t should be approximated as

$$E(Y_t) = E(\mu_t) \approx \exp(\mathbf{x}_t^T \boldsymbol{\beta})$$

so that the regression coefficients $\boldsymbol{\beta}$ can be interpreted as the proportional change in the marginal expectation of Y_t given a unit change in the regressor variable. D.1 is useful for interpretation.

D.2 Both positive and negative serial dependence should be possible in the model.

D.3 The estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ should be approximately orthogonal making their estimation easier (presumably because a two stage procedure could be used).

Condition D.3 is met for linear regression models with time series errors. However, for modeling count data, D.3 may be overly restrictive since the conditional mean and variance of Y_t are linked. Also with rapid computing now widely available requirement D.3 is no longer necessary.

In addition to these 3 criteria, we offer 3 more which fill out the modeling paradigm for count data. These are:

D.4 The ability to easily forecast with the model. This is often the primary goal in many time series applications.

D.5 A method for model fitting and inference should be reasonably straightforward to implement and control.

D.6 Diagnostic tools should be available for identification of a class of models, for the assessment of model adequacy, and for detection of outliers, etc.

1.3 Generalized State-Space Models

Linear state-space models and the associated Kalman recursions have had a profound impact on time series analysis and related areas. The techniques were originally developed in connection with the control of linear systems (for accounts of this subject see Davis and Vinter, 1985, and Hannan and Deistler, 1988). In recent years, non-linear state-space models have been developed to handle a wide range of situations not easily covered under the linear framework. In this subsection, we provide a brief overview of generalized state-space models.

A state-space model for a time series $\{Y_t, t = 1, 2, \ldots, \}$ consists of two equations referred to as the observation and state equations. Generalized state-space models can be loosely characterized as either "parameter driven" or "observation driven." The observation equation is the same for both models, but the state vectors of a parameter driven model evolve independently of the past history of the observation process, while the state vectors of an observation driven model depend on past observations.

The **observation equation** specifies the distribution of Y_t given a state variable S_t. For ease of presentation, we assume the state-variable is univariate, although it is often taken to be vector valued. Specifically, if $\mathbf{Y}^{(t)}$ and $\mathbf{S}^{(t)}$ denote the t–dimensional column vectors $\mathbf{Y}^{(t)} = (Y_1, Y_2, \ldots, Y_t)^T$ and $\mathbf{S}^{(t)} = (S_1, S_2, \ldots, S_t)^T$, respectively, then it is assumed that Y_t given $(S_t, \mathbf{S}^{(t-1)}, \mathbf{Y}^{(t-1)})$ is independent of $(\mathbf{S}^{(t-1)}, \mathbf{Y}^{(t-1)})$ with conditional probability density,

$$p(y_t|s_t, \mathbf{s}^{(t-1)}, \mathbf{y}^{t-1}) = p(y_t|s_t), \quad t = 1, 2, \ldots. \tag{1.1}$$

For the parameter-driven model, S_{t+1} given $(S_t, \mathbf{S}^{(t-1)}, \mathbf{Y}^{(t)})$ is assumed to be independent of $(\mathbf{S}^{(t-1)}, \mathbf{Y}^{(t)})$ with conditional density function,

$$p(s_{t+1}|s_t, \mathbf{s}^{(t-1)}, \mathbf{y}^{(t)}) := p(s_{t+1}|s_t), \quad t = 1, 2, \ldots. \tag{1.2}$$

This latter equation is known as the **state-equation** for the parameter driven model.

The joint density of the observation and state variables can be computed directly from (1.1)–(1.2) as

$$\begin{aligned} p(y_1, \ldots, y_n, s_1, \ldots, s_n) &= p(y_n|s_n, \mathbf{s}^{(n-1)}, \mathbf{y}^{(n-1)}) p(s_n, \mathbf{s}^{(n-1)}, \mathbf{y}^{(n-1)}) \\ &= p(y_n|s_n) p(s_n|\mathbf{s}^{(n-1)}, \mathbf{y}^{(n-1)}) p(\mathbf{y}^{(n-1)}, \mathbf{s}^{(n-1)}) \\ &= p(y_n|s_n) p(s_n|s_{n-1}) p(\mathbf{y}^{(n-1)}, \mathbf{s}^{(n-1)}) \\ &= \cdots \\ &= \left(\textstyle\prod_{j=1}^{n} p(y_j|s_j)\right) \left(\textstyle\prod_{j=2}^{n} p(s_j|s_{j-1})\right) p_1(s_1), \end{aligned} \tag{1.3}$$

and since (1.2) implies that $\{S_t\}$ is Markov

$$p(y_1, \ldots, y_n|s_1, \ldots, s_n) = \left(\prod_{j=1}^{n} p(y_j|s_j)\right). \tag{1.4}$$

We conclude that $Y_1, \ldots, Y_n$ are conditionally independent given the state variables $S_1, \ldots, S_n$, so that the dependence structure of $\{Y_t\}$ is inherited from that of the state-process $\{S_t\}$. The sequence

of state-variables $\{S_t\}$ is often referred to as the **hidden** or **latent** generating process associated with the observed process.

In an observation-driven model specification, it is again assumed that Y_t, conditional on the vector $(S_t, \mathbf{S}^{(t-1)}, \mathbf{Y}^{(t-1)})$, is independent of $(\mathbf{S}^{(t-1)}, \mathbf{Y}^{(t-1)})$. The model is specified by the conditional densities

$$p(y_t|\mathbf{s}^{(t)}, \mathbf{y}^{(t-1)}) = p(y_t|s_t), \quad t = 1, 2, \ldots, \tag{1.5}$$

$$p(s_{t+1}|\mathbf{y}^{(t)}) = p_{S_{t+1}|\mathbf{Y}_{(t)}}(s_{t+1}|\mathbf{y}^{(t)}), \quad t = 0, 1, \ldots, \tag{1.6}$$

where $p(s_1|y^{(0)}) := p_1(s_1)$ for some prespecified initial density $p_1(s_1)$.

The advantage of the observation driven state equation (1.6) is that the forecast density $p(y_{t+1}|\mathbf{y}^t)$ is easy to compute from the relation

$$p(y_{t+1}|\mathbf{y}^{(t)}) = \int p(y_{t+1}|s_{t+1})p(s_{t+1}|\mathbf{y}^{(t)})\, d\mu(s_t). \tag{1.7}$$

In other words computing the best predictor of Y_{t+1} in terms of $\mathbf{Y}^{(t)}$ and calculating the joint density function of $\mathbf{y}_n$ given by

$$p(y_1, \ldots, y_n) = \prod_{t=1}^{n} p(y_t|\mathbf{y}^{(t-1)}) \tag{1.8}$$

are relatively simple tasks. Calculation of the joint density function is particularly important for estimation and making inferences about the parameters of the model. On the other hand, for a parameter driven model, computation of the forecast density function and hence the joint density function, is difficult requiring recursive updating of the densities $p(s_t|\mathbf{y}^{(t)})$ and $p(s_{t+1}|\mathbf{y}^{(t)})$ via Bayes's theorem (see Brockwell and Davis (1996)).

While the observation-driven model formulation has a decided edge for forecasting future values of the series and for calculating the joint density, the evolutionary properties of the time series are more difficult to characterize than for the parameter-driven model. Specifically, ergodic and asymptotic stationarity of a time series based on a observation-driven model can be quite difficult to establish and this is the subject of ongoing research. However, for a process based on a parameter-driven model, such properties are typically inherited by those assumed for the underlying state-variables.

1.4 Two Examples

To illustrate the issues, theory and methods we will use the following two examples throughout our discussion.

1.4.1 Asthma Presentations at a Sydney Hospital

The data arose from a single hospital (at Cambelltown) as part of a larger (ongoing) study into the relationship between atmospheric pollution and the number of asthma cases presenting themselves to various emergency departments in local hospitals in the South West region of Sydney, Australia. It is not our purpose here to examine this issue since the data required to do this properly is not yet fully compiled. In addition, for much of the record that we examine, measurements on other variables, such as pollen counts, which could influence the asthma attacks that people experience, are not available. However, the analysis we provide here would be required as a first step in constructing a model relating the putative causes and the outcome counts. Regional measurements

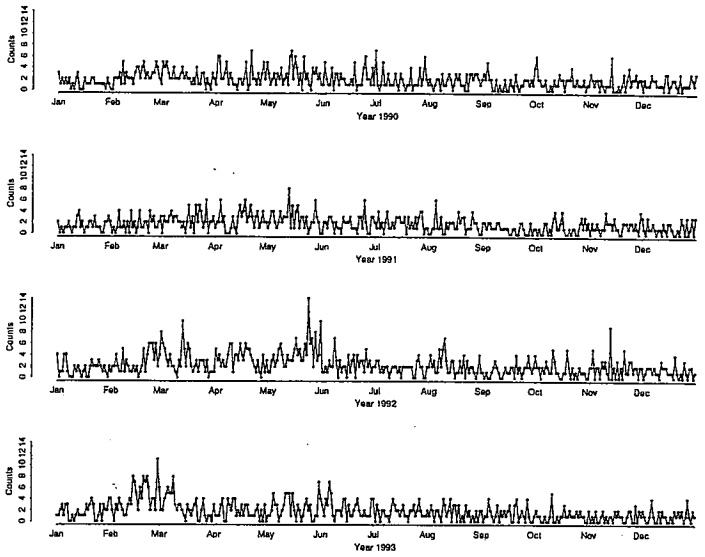

Figure 1: Asthma presentations at a Sydney hospital

are also required because the weather patterns in Sydney are such as to move pollution around substantially from day to day.

Figure 1 shows the daily number of asthma presentations from January 1, 1990 – December 31, 1993. It is apparent from this figure that there is some form of seasonal pattern with higher activity occurring in the Fall (March - May). There is also a suggestion of an increasing trend in time. Note that this is a distinct possibility regardless of whether asthma is more prevalent in the region simply because it is an area which has a young and growing population. Of interest would be to test the significance of any trend, suitably adjusted for serial dependence. There are also patterns of departure from a seasonal cycle in the data (e.g. in 1992) which exhibit clear signs of positive serial dependence.

Not apparent, but consistent with behavioral aspects of the population, is the possibility of a day of the week effect (which also could be confounded with weekly pattern in pollution). General practitioners (private physicians) are less available on Sundays and people tend to rely more on the emergency departments of hospitals should an asthma attack occur. As it turns out that the day of the week effect can be characterized by being similar on Tuesday through Saturday and elevated incidence rate occurring on Sunday and Monday with roughly similar size effects on these two days. The Monday rise might be explained in terms of patients who hold off through Sunday hoping for improvement but who then require hospital attention on Monday. In any case testing the significance of day of week effects is important in daily data of these type.

1.4.2 Polio Incidence

The second data set that we will consider consists of the monthly number of cases of poliomyelitis in the U.S. for the year 1970–1983 as reported by the Center for Disease Control. This data, which was originally presented in Zeger (1988), has become a standard example in the field. The data, which is graphed in Figure 2 reveals some seasonality and the possibility of a slight decreasing trend. The main objective in modelling this data is the detection of a decreasing trend. Although some

researchers have considered the observed count of 14 for November, 1972 to be an outlier, there is no corroborating evidence—we did not remove it in order to compare with existing analyses reported here which also did not remove or adjust it. Chan and Ledolter (1995) found a slight change in the slope and intercept terms of the model when the analysis is adjusted for the outlier.

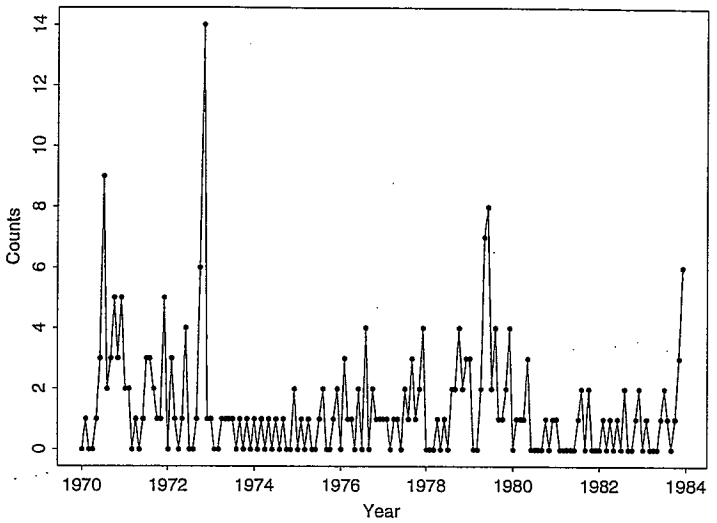

Figure 2: Monthly number of cases of poliomyelitis in the U.S. (1970-1983)

1.5 Outline of Paper

The focus of this paper is on the similarities and differences in model building and interpretation for the two types of conditional mean processes, the parameter driven and observation driven specifications. We will concentrate exclusively on the case where the observations, conditional upon a specified mean process, are independent Poisson observations. This class of models is useful in a wide range of applications in their own right. Focusing on the Poisson case allows the exposition of the main ideas without the distraction of the more general class of exponential family models.

Section 2 is concerned with defining the parameter driven model in terms of a latent process specification. The probabilistic properties of such models are developed. Inferential methods are reviewed and illustrated on the time series of asthma counts and the polio data. Section 3 provides a parallel exposition for observation driven models along with a comparison between them on the same examples.

2 Parameter Driven Models

In this section, we consider a class of parameter-driven or latent process models for modeling time series of counts. We will focus on the situation when the 'observation equation' is governed by a Poisson distribution. Denote the time series of counts by $Y_1, \ldots, Y_n$ and suppose that for each t, $\mathbf{x}_t$ is a p-dimensional regression vector whose first component is 1. It will also be convenient to introduce the notation $\mathbf{Y}^{(t)} = (Y_1, \ldots, Y_t)^T$, $X^{(t)} = (\mathbf{x}_1, \ldots, \mathbf{x}_t)$, and $\epsilon^{(t)} = (\epsilon_1, \ldots, \epsilon_t)^T$, where

$\{\epsilon_t\}$ is the latent process. In this setting, the state-variable S_t described in Section 1.3, can be set to either the multivariate vector consisting of $(\mathbf{x}_t^T, \epsilon_t)^T$ or just ϵ_t.

The conditional distribution of Y_t given $\mathbf{x}_t$ and ϵ_t corresponding to the observation equation of (1.1) is assumed to be Poisson with mean $u_t = \epsilon_t \exp\{\mathbf{x}_t^T\boldsymbol{\beta}\}$ denoted by,

$$Y_t|\epsilon_t, \mathbf{x}_t \sim \mathcal{P}(\epsilon_t \exp\{\mathbf{x}_t^T\boldsymbol{\beta}\}), \tag{2.1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$. The analogue of the state-equation in this context is a specification of the distributional properties of the latent process $\{\epsilon_t\}$. Unlike the standard state-space formulation considered in Section 1.3, it is not necessary for the ϵ_t to have a Markov structure. At this point, we only assume that $\{\epsilon_t\}$ is a non-negative strictly stationary time series with mean 1 and autocovariance function (ACVF)

$$\gamma_\epsilon(h) = E(\epsilon_{t+h} - 1)(\epsilon_t - 1). \tag{2.2}$$

A Markov structure could be imposed on the state-variables by replacing the ϵ_t with high-dimensional random vectors consisting of the lagged components of the ϵ_t. However, for our purposes, this reformulation offers no real advantage.

The assumption of non-negativity of the ϵ_t is clear in order to ensure that the conditional mean of Y_t is non-negative. The condition that $E(\epsilon_t) = 1$ is imposed for identifiability reasons; otherwise, if $c = E(\epsilon_t) \neq 1$, then c can be absorbed into the intercept term in the exponent of μ_t. (That is one would replace ϵ_t with ϵ_t/c and β_1 with $\beta_1 + \ln c$.)

In order to meet the non-negativity constraint on the ϵ_t, it is often convenient to model the logarithms of the ϵ_t. Letting $\delta_t = \ln \epsilon_t$, then the conditional mean of Y_t can be written as

$$u_t = \exp\{\mathbf{x}_t^T\boldsymbol{\beta} + \delta_t\}.$$

Of course, in order for the corresponding ϵ_t to have mean 1, we must assume $Ee^{\delta_t} = 1$. Unless the $\{\delta_t\}$ is a stationary Gaussian process, there is not an explicit relationship between the ACVF's of $\{\epsilon_t\}$ and $\{\delta_t\}$. In the case when $\{\epsilon_t\}$ is a stationary log-normal process, i.e., $\{\delta_t\}$ is a stationary Gaussian process with ACVF $\gamma_\delta(\cdot)$, then there is a nice connection between the ACVF's. First, in order to satisfy the identifiability requirement that $E(\epsilon_t) = E[\exp(\delta_t)] = 1$ it is required that $\delta_t \sim N(-\sigma_\delta^2/2, \sigma_\delta^2)$ $(\sigma_\delta^2 := \gamma_\delta(0))$, so that the mean is half of the variance. Then, with this choice of mean and variance in the lognormal distribution, $\gamma_\epsilon(h) = E(\exp\{\delta_{t+h} + \delta_t\} - 1) = e^{\gamma_\delta(h)} - 1$ for all h.

2.1 Means, Variances, and Autocovariances of Y_t

In this section various key facts about the moments of the observed count process, Y_t, are derived and relationships between the first and second moments of the latent process, ϵ_t, are provided. While most of these results are available elsewhere (Zeger (1988) for example) it is useful to have these collected together for easy reference. Throughout, expectations, variances, and covariances are conditional upon the regressors $\mathbf{x}_t$ (and this will not be explicitly noted) but not on the latent process unless otherwise indicated in the usual way. Recall that $E(\epsilon_t) = 1$.

Mean of Y_t:

$$\mu_t = E(Y_t) = E(E(Y_t|\epsilon_t)) = \exp(\mathbf{x}_t^T\boldsymbol{\beta}).$$

Variance of Y_t:

$$\mathrm{Var}(Y_t) = E(\mathrm{Var}(Y_t|\epsilon_t)) + \mathrm{Var}(E(Y_t|\epsilon_t)) = \mu_t + \sigma_\epsilon^2\mu_t^2.$$

Autocovariance function of Y_t:

$$\mathrm{Cov}(Y_{t+h}, Y_t) = \mu_t \mu_{t+h}(E(\epsilon_t \epsilon_{t+h}) - 1) = \mu_t \mu_{t+h} \gamma_\epsilon(h).$$

Autocorrelation function of Y_t:

$$\begin{aligned} \mathrm{Cor}(Y_s, Y_t) &= \frac{\mu_s \mu_t \gamma_\epsilon(s-t)}{\sqrt{[\mu_s + \mu_s^2 \sigma_\epsilon^2][\mu_t + \mu_t^2 \sigma_\epsilon^2]}} \\ &= \frac{\rho_\epsilon(s-t)}{\sqrt{[1 + (\sigma_\epsilon^2 \mu_s)^{-1}][1 + (\sigma_\epsilon^2 \mu_t)^{-1}]}} \end{aligned}$$

and this is not free of the regressors $\mathbf{x}_t$, as is to be expected. In the case when there are no regression terms other than a constant, i.e., $p = 1$, the process $\{Y_t\}$ is then stationary with autocorrelation function (ACF) given by

$$\rho_Y(h) := \mathrm{Cor}(Y_{t+h}, Y_t) = \frac{\gamma_\epsilon(h)}{\mu^{-1} + \sigma_\epsilon^2},$$

where $\mu = e^{\beta_1}$. Since $\mu > 0$ we see that

$$|\rho_Y(h)| \leq |\rho_\epsilon(h)|,$$

where $\rho_\epsilon(h) =: \mathrm{Cor}(\epsilon_{t+h}, \epsilon_t)$. To illustrate this last observation, consider Figure 3 in which the ACF for the $\{Y_t\}$ process is shown along with that for the latent process. Clearly, even when the mean of Y_t is stationary, the ACF of the observed count process tends to underestimate that of the latent process. Because of this methods are required to estimate the underlying correlation and to test if it is zero or not. Such methods also need to be applicable when the regression terms are present.

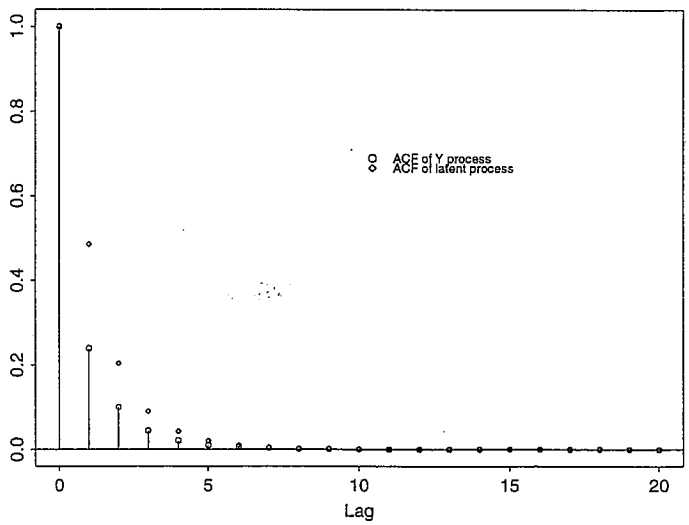

Figure 3: Autocorrelation functions of the $\{Y_t\}$ and latent processes

2.2 Preliminary Estimates and Diagnostics

In a linear model with time series errors, e.g.,

$$Y_t = \mathbf{x}_t^T \boldsymbol{\beta} + \epsilon_t,$$

the first step in fitting such model, is to determine the autocovariance structure of the time series of errors $\{\epsilon_t\}$. Assuming that $\{\epsilon_t\}$ is a linear process such as an ARMA, the parameter $\boldsymbol{\beta}$ is estimated using ordinary least squares (OLS) by regressing the data vector $(Y_1, \ldots, Y_n)^T$ onto the $\mathbf{x}_t$. While this estimate ignores the dependence structure of the $\{\epsilon_t\}$, the OLS estimate has the same asymptotic efficiency as the MLE of $\boldsymbol{\beta}$ under a wide class of models for the ϵ_t process (see for example Hannan (1970)). The asymptotic covariance matrix of the OLS (and MLE) estimate does depend on the covariance structure of the ϵ_t. Once a consistent estimator of $\boldsymbol{\beta}$ has been found, then the ACVF of the ϵ_t can be consistently estimated from the sample ACVF of the residuals defined by, $\hat{\epsilon}_t = Y_t - \mathbf{x}_t^T \hat{\boldsymbol{\beta}}_{OLS}$. A model is then selected for the regression parameter $\boldsymbol{\beta}$ and the parameters of the model for the ϵ_t can be re-estimated using MLE.

In this section we consider carrying out the same procedure applied to our parameter-driven model. The first step is to estimate the $\boldsymbol{\beta}$ vector using generalized linear models (GLM) or Poisson regression.

The GLM estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is obtained by maximizing the function

$$l(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{t=1}^{n} e^{\mathbf{x}_t^T \boldsymbol{\beta}} + \frac{1}{n}\sum_{t=1}^{n} Y_t \mathbf{x}_t^T \boldsymbol{\beta}.$$

Note that $l(\boldsymbol{\beta})$ is the log-likelihood (apart from a constant) corresponding to a Poisson regression model without a latent process. In order to derive the asymptotic behavior of $\hat{\boldsymbol{\beta}}$ we assume there exists a sequence of nonsingular matrices M_n such that the regressors obey the following conditions:

$$M_n \left(\sum_{t=1}^{n} \mathbf{x}_t \mathbf{x}_t^T e^{\mathbf{x}_t^T \boldsymbol{\beta}} \right) M_n^T \to \Omega_I(\boldsymbol{\beta}), \tag{2.3}$$

$$M_n \left(\sum_{t=1}^{n} \mathbf{x}_t \mathbf{x}_{t+h}^T e^{(\mathbf{x}_t^T + \mathbf{x}_{t+h}^T)\boldsymbol{\beta}} \right) M_n^T \to \Omega_h(\boldsymbol{\beta}) \text{ (uniformly in } h), \tag{2.4}$$

$$\max_{1 \le t \le n} |M_n^T \mathbf{x}_t| e^{\mathbf{x}_t^T \boldsymbol{\beta}} \to 0, \tag{2.5}$$

and for $h \le 0$,

$$M_n \sum_{t=1}^{1-h} \mathbf{x}_t \mathbf{x}_{t+h}^T e^{(\mathbf{x}_t^T + \mathbf{x}_{t+h}^T)\boldsymbol{\beta}} M_n^T \to 0, \tag{2.6}$$

and is uniformly bounded in h as $n \to \infty$. Similarly, for $h > 0$,

$$M_n \sum_{t=n-h}^{n} \mathbf{x}_t \mathbf{x}_{t+h}^T e^{(\mathbf{x}_t^T + \mathbf{x}_{t+h}^T)\boldsymbol{\beta}} M_n^T \to 0 \tag{2.7}$$

and is uniformly bounded in h as $n \to \infty$. Following the theorem below, we show that these conditions are met for a wide range of regression variables. Before stating the theorem, define

$$\begin{aligned} \Omega_n &:= \operatorname{Cov}(M_n \sum_{t=1}^{n} \mathbf{x}_t (Y_t - \mu_t)) \\ &= M_n \sum_{t=1}^{n} \sum_{s=1}^{n} \mathbf{x}_t E\left[(Y_t - \mu_t)(Y_s - \mu_s)\right] \mathbf{x}_t^T M_n^T. \end{aligned}$$

Theorem 2.1 *Let $\hat{\beta}$ be the GLM estimate of β obtained by minimizing $l(\beta)$ (that is, assuming that the latent process $\{\epsilon_t\}$ is not present in the model for the observed mean) for the parameter-driven model specified in (1.1)–(1.4). Further assume that the $\{\mathbf{x}_t\}$ satisfy (2.3)–(2.7), and $\sum_{h=0}^{\infty} |\gamma_\epsilon(h)| < \infty$. Then,*

$$\Omega_n \rightarrow \Omega_I(\beta) + \Omega_{II}(\beta) \tag{2.8}$$

and

$$\hat{\beta} \rightarrow \beta, \tag{2.9}$$

where

$$\Omega_{II}(\beta) = \sum_{h=-\infty}^{\infty} \Omega_h(\beta)\gamma_\epsilon(h).$$

Moreover, if

$$M_n \sum_{t=1}^{n} \mathbf{x}_t e^{\mathbf{x}_t^T \beta}(\epsilon_t - 1) \xrightarrow{d} N(\mathbf{0}, \Omega_{II}(\beta)), \tag{2.10}$$

then

$$M_n^{-1}(\hat{\beta} - \beta) \rightarrow N(\mathbf{0}, \Omega_I^{-1}(\beta) + \Omega_I^{-1}(\beta)\Omega_{II}(\beta)\Omega_I^{-1}(\beta))$$

as $n \rightarrow \infty$.

REMARK 2.1. The matrix, $\Omega_I^{-1}(\beta)$, is the asymptotic covariance matrix associated with the usual GLM estimate, while $\Omega_I^{-1}(\beta)\Omega_{II}(\beta)\Omega_I^{-1}(\beta)$ is the additional contribution to the asymptotic covariance due to the existence of the latent process. Clearly if the latent process has small covariance then this second term will not contribute very much.

REMARK 2.2. In the following subsection, conditions under which the various covariance matrices converge to their asymptotic limits are described and in some cases the form of the limit is given explicitly in terms of integrals.

REMARK 2.3. The convergence in (2.10) holds under a variety of conditions on the latent process. For example, if $\{\epsilon_t\}$ is strongly mixing with a suitable rate or if $\{\delta_t = \ln \epsilon_t\}$ is a linear process, then the central limit theorem in (2.10) holds.

Brannas and Johansson (1994) state that results of Judge (1985) can be used to establish the consistency and asymptotic normality of the GLM estimates even when autocorrelation is present. The form of the covariance matrix given above shows explicitly how autocorrelation inflates the true asymptotic covariance matrix.

2.2.1 Calculating the Asymptotic Covariance Matrix

Conditions (2.3)–(2.7) hold for a variety of regression functions including:

1. Trend type terms in which the regression function depends on n and t and has the form

$$\mathbf{x}_t = \mathbf{f}(t/n),$$

for some vector-valued continuous function $\mathbf{f}(\cdot)$ defined on the unit interval $[0, 1]$. For example $\mathbf{f}$ could be a vector of polynomials. In particular, when $\mathbf{f}^T(x) = (1, x)$ this would specify a linear trend term in the regression. On the other hand, the choice of regression variables $\mathbf{x}_t^T = (1, t)$ would not give rise to a consistent estimate of β for $\beta_2 \leq 0$. Two attractive

features about using regression variables of the form $\mathbf{x}_t = \mathbf{f}(t/n)$ are that the asymptotic covariance matrix can be computed in closed form and the normalizing matrix M_n is given by $n^{-1/2}I_p$, where I_p is the identity matrix. To see this, note that

$$\begin{aligned}
\Omega_n &:= \frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n}\mathbf{x}_t\mathbf{x}_s^T E\left[(Y_t-\mu_t)(Y_s-\mu_s)\right] \\
&= \frac{1}{n}\sum_{t=1}^{n}\mathbf{x}_t\mathbf{x}_t^T\mu_t + \frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n}\mathbf{x}_t\mathbf{x}_s^T\mu_t\mu_s\gamma_\epsilon(s-t) \\
&= \frac{1}{n}\sum_{t=1}^{n}\mathbf{f}(t/n)\mathbf{f}^T(t/n)e^{\mathbf{f}^T(t/n)\boldsymbol{\beta}} + \frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n}\mathbf{f}(t/n)\mathbf{f}^T(s/n)e^{(\mathbf{f}^T(t/n)+\mathbf{f}^T(s/n))\boldsymbol{\beta}}\gamma_\epsilon(s-t) \\
&\to \Omega_I(\boldsymbol{\beta}) + \Omega_{II}(\boldsymbol{\beta}),
\end{aligned}$$

where

$$\Omega_I(\boldsymbol{\beta}) = \int_0^1 \mathbf{f}(x)\mathbf{f}^T(x)e^{\mathbf{f}^T(x)\boldsymbol{\beta}}dx \quad \text{and} \quad \Omega_{II}(\boldsymbol{\beta}) = \left(\int_0^1 \mathbf{f}(x)\mathbf{f}^T(x)e^{2\mathbf{f}^T(x)\boldsymbol{\beta}}dx\right)\sum_{h=-\infty}^{\infty}\gamma_\epsilon(h).$$

2. Harmonic terms to specify annual effects or day of the week effects. For example $x_t = \cos(2\pi t/12)$. This is an example of an asymptotically stationary process.

3. Stationary processes as might arise from a seasonally adjusted temperature series. See Campbell (1994) for an example of this.

2.2.2 Application to the Polio Data

We now apply the results of Theorem 2.1 to the Polio data example from Zeger (1988). Here we use the same regression variables as in Zeger (1988) consisting of an intercept term, a linear trend, and harmonics at periods of 6 and 12 months. Specifically,

$$\mathbf{x}_t = (1, t'/1000, \cos(2\pi t'/12), \sin(2\pi t'/12), \cos(2\pi t'/6), \sin(2\pi t'/6))^T.$$

where $t' = (t-73)$ to locate the intercept term at January, 1976 as in Zeger's analysis.

	Zeger		GLM Fit		Asym	Simulations	
	$\hat{\beta}_Z$	s.e.$(\hat{\beta}_Z)$	$\hat{\beta}_{GLM}$	s.e.$(\hat{\beta}_{GLM})$	s.e.$(\hat{\beta}_{GLM})$	mean $\hat{\beta}_{GLM}$	s.d. $\hat{\beta}_{GLM}$
Intercept	0.17	0.13	0.207	0.075	.205	.150	0.213
Trend $\times 10^{-3}$	-4.35	2.68	-4.799	1.403	4.115	-4.887	3.937
$\cos(2\pi t/12)$	-0.11	0.16	-0.149	0.097	0.157	-0.145	0.144
$\sin(2\pi t/12)$	-0.48	0.17	-0.532	0.109	0.168	-0.531	0.168
$\cos(4\pi t/12)$	0.20	0.14	0.169	0.098	0.122	0.167	0.123
$\sin(4\pi t/12)$	-0.41	0.14	-0.432	0.101	0.125	-0.440	0.125

The asymptotic standard errors for the GLM estimates given in the above theorem are estimated using the values of $\sigma_\epsilon^2 = 0.77$ and $\hat{\rho}_\epsilon(h) = (0.77)^h$ reported in Zeger (1988, Table 3). These were obtained from the formula

$$\Omega_I = \frac{1}{n}\sum_{t=1}^{n}\mathbf{f}(t/n)\mathbf{f}^T(t/n)e^{\mathbf{f}^T(t/n)\boldsymbol{\beta}} \quad \text{and } \Omega_{II} = \frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n}\mathbf{f}(t/n)\mathbf{f}^T(s/n)e^{(\mathbf{f}^T(t/n)+\mathbf{f}^T(s/n))\boldsymbol{\beta}}\hat{\gamma}_\epsilon(s-t).$$

Use of the correct standard errors for the trend term would lead to the conclusion that the trend is not significant whereas use of the standard errors produced by the GLM analysis would lead to declaring the trend to be significant.

The final two columns of the above table report the results of 1,000 simulations of time series of length $n = 168$ using the GLM fitted values, $\hat{\beta}_{GLM}$, as the true values. The latent process was assumed to be a lognormal autoregression with $\phi = 0.82$ and $\sigma_\delta^2 = 0.57$. The mean of this autoregression was chosen as $-\sigma_\delta^2/2 = -0.285$ in order to satisfy the conditions of the above theorem. The average over the 1,000 simulated values of $\hat{\beta}_1$ is observed as 0.150 which is significantly biased downwards from the true value of .207 used in the simulations. The other parameters appear to be estimated without substantial bias. The standard deviation of the GLM estimates observed over the 1,000 replications are reported in the last column of the above table. These are in good agreement with the standard errors obtained from the asymptotic theory (column 6).

2.3 Testing for the Existence of a Latent Process

Prior to the estimation of autocovariances it is reasonable to test for the existence of a latent process. Brannas and Johansson (1994) review the following statistic

$$S = \frac{\sum_{t=1}^{n}[(Y_t - \hat{\mu}_t)^2 - Y_t]}{[2\sum_{t=1}^{n}\hat{\mu}_t^2]^{1/2}}$$

derived by several authors and based on a local alternative hypothesis or the Lagrange multiplier test of the Poisson distribution against a negative binomial or more general Katz distribution. A variant, introduced by Dean and Lawless (1989) "to improve on the small sample performance of the test", also considered by them is

$$S_a = \frac{\sum_{t=1}^{n}[(Y_t - \hat{\mu}_t)^2 - Y_t + \hat{h}_t\hat{\mu}_t]}{[2\sum_{t=1}^{n}\hat{\mu}_t^2]^{1/2}},$$

where h_t is the tth diagonal element of the "hat" matrix. The "hat" matrix for generalized linear models extends that for linear regression and is defined in Fahrmeir and Tutz (1994, p.127), for example, as $H = \Lambda^{1/2}X(X^T\Lambda X)^{-1}X^T\Lambda^{1/2}$, where $\Lambda = \text{diag}(\mu_1, \cdots, \mu_n)$ and $X = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)^T$ is the design matrix. Note that use of the $\hat{h}_t$ adjusts for the first part, $\hat{\Omega}_I = (X^T\Lambda X)$, of the asymptotic covariance matrix in Theorem 1. Under the hypothesis of no latent process this is appropriate. Later, when we examine the bias in autocovariance estimates when there is a latent process, additional adjustments are required to account for the complete asymptotic variance of $\hat{\beta}$.

Both of these statistics are asymptotically distributed as a $N(0,1)$ variate under the null hypothesis of no latent process. They are used in a one-sided test. Monte Carlo work reviewed in Brannas and Johansson and corroborated by us, suggests that S_a has better size properties in small samples. For this reason, we will consider S_a from here on.

We also introduce an alternative test specifically designed for overdispersion due to the existence of a latent process in a Poisson observation process. Since this test uses higher moment properties of Poisson observations it was considered as a possibly more powerful test statistic than S_a. Also it might be more appropriate for the lagged regression models introduced in Section 3 since for these the distributional theory of S_a is not clearly appropriate. Under the null hypothesis that there is no latent process (i.e. $\epsilon_t \equiv 1$) the Pearson residuals

$$e_t = \frac{Y_t - \hat{\mu}_t}{\sqrt{\hat{\mu}_t}}$$

have approximately zero mean and unit variance. Hence the statistic

$$Q = (\frac{1}{n}\sum_{t=1}^{n} e_t^2 - 1)/\hat{\sigma}_Q,$$

where

$$\hat{\sigma}_Q^2 = \frac{1}{n}(\frac{1}{n}\sum_{t=1}^{n}\frac{1}{\hat{\mu}_t} + 2),$$

could be used to test for a latent process. The expression for $\hat{\sigma}_Q^2$ is easily derived using the fact that a Poisson random variable Y_t with mean μ_t has fourth central moment $E(Y_t - \mu_t)^4 = \mu_t + 3\mu_t^2$. Under the hypothesis that the variance of the latent process is zero

$$Q \sim N(0,1)$$

approximately.

Using 1000 replications of a time series of length $n = 168$ obtained from simulating independent Poisson variates (i.e. with no latent process present) with mean $\mu_t = \hat{\mu}_t$, the GLM fit to the polio data considered in Zeger (1988), gave simulated type I errors for Q which are severely lower than the nominal values. The observed mean and standard deviation of Q are -0.23 and 0.788, respectively, explaining the low coverage type I errors rates observed. To adjust for the negative bias alternate estimates of residuals which adjust for the effect of fitted values $\hat{\mu}_t$ could be used. For example one could use the divisor $n - p$ instead of n in the numerator of Q. The resulting statistic had appreciably better performance than Q but was still on the conservative side. A second reasonably simple approach would be to use standardised Pearson residuals. Standardised Pearson residuals are

$$\tilde{e}_t = e_t/(1 - h_t)^{0.5},$$

where h_t is the tth value of the "hat" matrix. The mean value of these h_t values is p (the dimension of β). Using these standardised residuals we define

$$\tilde{Q} = (\frac{1}{n}\sum_{t=1}^{n} \tilde{e}_t^2 - 1)/\hat{\sigma}_Q.$$

With this definition we get based on the same 1000 replications as used for Q above, mean and standard deviation of $\tilde{Q}$ as 0.011 and 0.826, respectively, and clearly improved, although still low, Type I errors as

α	.100	.050	.025	.010
Empirical $P(\tilde{Q} > z_{1-\alpha})$	.073	.037	.022	.004

with corresponding significance points

α	.100	.050	.025	.010
$z^*_{1-\alpha}$ s.t. $P(\tilde{Q} > z^*_{1-\alpha})$	1.10	1.48	1.89	2.21

On this basis, $\tilde{Q}$ is preferred to the unadjusted version Q.

We next compare the size and size adjusted power of the two statistics $\tilde{Q}$ and S_a. First the size properties are as follows for a simulation of 1000 replicates assuming no latent process is present:

Case 1: Linear Regression $1 + 1t/100$ for $t = 1, \ldots, 100$.

α	.100	.050	.025	.010
Empirical $P(\tilde{Q} > z_{1-\alpha})$	.089	.048	.026	.003
Empirical $P(S_a > z_{1-\alpha})$	.085	.045	.027	.011

Case 2: Cosine Regression 1 $+\cos(2\pi t/12)$ for $t = 1, \ldots, 100$.

α	.100	.050	.025	.010
Empirical $P(\tilde{Q} > z_{1-\alpha})$	.077	.038	.021	.013
Empirical $P(S_a > z_{1-\alpha})$	.099	.056	.025	.009

In addition the power of the test to detect departures from the null hypothesis were investigated using 1000 replicates with the same regression models. The latent process was generated using a lognormal distribution. The latent process had variance $\sigma_\epsilon^2 = 0.05$ chosen to give a small deviation from the null hypothesis. The autocovariance was simulated using an autoregressive process with $\phi = 0$ and $\phi = 0.9$. The results, again based on 1000 replications are for a size 0.05 test using empirical significance points obtained under the null hypothesis.

	Linear Regression 1 $+ 1t/100$		Cosine Regression 1 $+ 1\cos(2\pi t/12)$	
	$\phi = 0$	$\phi = 0.9$	$\phi = 0$	$\phi = 0.9$
Power of $\tilde{Q}$	.460	.306	.290	.230
Power of S_a	.491	.316	.410	.342

On the basis of this limited simulation it would appear as though the S_a statistic has better type I error rates (i.e. closer to normal distribution) and larger power especially for a regressor which is seasonally varying. Note also that if the latent process has positive autocorrelation ($\phi = 0.9$) the power of S_a and $\tilde{Q}$ are reduced relative to the case of a white noise latent process ($\phi = 0$). Further research is required to demonstrate that S_a is preferred in all plausible circumstances.

2.4 Estimating the Variance of the Latent Process

We now assume that the test proposed in the last section rejects the hypothesis of no latent process. Zeger proposed the estimate

$$\hat{\sigma}^2_{\epsilon,Z} = \frac{\sum_{t=1}^n [(Y_t - \hat{\mu}_t)^2 - \hat{\mu}_t]}{\sum_{t=1}^n \hat{\mu}_t^2}$$

of σ_ϵ^2 which is approximately unbiased for σ_ϵ^2 and would be exactly unbiased if the $\hat{\mu}_t$ were replaced by the true values μ_t. Brannas and Johansson (1994) suggested the estimate

$$\hat{\sigma}^2_{OLS} = \frac{\sum_{t=1}^n \hat{\mu}_t^2 [(Y_t - \hat{\mu}_t)^2 - \hat{\mu}_t]}{\sum_{t=1}^n \hat{\mu}_t^4}$$

which is derived by using ordinary least squares (OLS) regression of $(Y_t - \hat{\mu}_t)^2 - \hat{\mu}_t$ on its approximate expected value of $\sigma_\epsilon^2 \hat{\mu}_t^2$. (Recall $\mathrm{Var}(Y_t) = \mu_t + \sigma_\epsilon^2 \mu_t^2$.)

The above estimates are not necessarily optimal in any sense other than being approximately unbiased. More generally, consider weighted estimates of the form:

$$\hat{\sigma}^2_{\epsilon,W} = (\sum_{t=1}^n W_t)^{-1} \sum_{t=1}^n W_t E_t,$$

where

$$E_t = \hat{\mu}_t^{-1}[\tilde{e}_t^2 - 1]$$

and $\tilde{e}_t$ is the adjusted Pearson residual introduced earlier. Note that because $E(\hat{\sigma}^2_{\epsilon,W}) \approx \sigma^2_\epsilon$ the weighted estimate is approximately unbiased and would be exactly so if $\hat{\mu}_t$ was replaced by μ_t. Zeger's estimate corresponds to choosing weights $W_t = \hat{\mu}_t^2$ and Brannas and Johansson's to choosing $W_t = \hat{\mu}_t^4$. Note also that these weighted estimates are not guaranteed to be positive. However, it is unlikely that a negative estimate will be produced from these methods if the test of the previous subsection is applied to determine if the latent process is present.

The optimal weights for minimizing the variance of $\hat{\sigma}^2_{\epsilon,W}$ are given by

$$W_t^* = 1/\mathrm{Var}(E_t).$$

Calculation of the variances required for this are complicated since they depend on moments up to order 4 for the latent process when it is a sequence of independent random variables. For latent processes with autocorrelation the calculation is further complicated. In addition these higher moments and autocovariances must be estimated.

Under the assumption that the latent process is a sequence of independent and identically distributed random variables, the optimal weights are

$$W_t^* = 1/\mathrm{Var}(E_t) \approx \mu_t^4/B_t,$$

where

$$B_t = E[(Y_t - \mu_t)^4] - (\mu_t + \sigma^2_\epsilon \mu_t^2)^2.$$

Calculation of these optimal weights would require an iterative approach starting with an initial estimate of γ_ϵ.

Note that the variance of the optimally weighted estimator is

$$\mathrm{Var}(\hat{\sigma}^2_{\epsilon,W}) \approx \frac{1}{(\sum_{t=1}^n \mu_t^4/B_t)}$$

and that of Zeger's estimator is

$$\mathrm{Var}(\hat{\sigma}^2_{\epsilon,Z}) \approx \frac{\sum_{t=1}^n B_t}{(\sum_{t=1}^n \mu_t^2)^2}.$$

For the polio data, using the GLM fit to obtain $\hat{\mu}_t$ and using the value of $\gamma = 1.77$ the values of these variances are approximately $\mathrm{Var}(\hat{\sigma}^2_{\epsilon,W}) \approx .46^2$ and $\mathrm{Var}(\hat{\sigma}^2_{\epsilon,Z}) \approx .53^2$ indicating a modest improvement in estimation of variance using the optimal weighting based on the assumption (incorrect in this case) that the latent process is independent.

In the non-independent case, the calculation of optimal weights will be complicated by their dependence on unknown covariances. From now on we will use the "optimal" weights derived above under the assumption of independence and evaluate their performance in the correlated case using simulations. To implement the above optimal weighting scheme an initial estimate of the variance is required. One possibility is to use the weights based on the assumption that the latent process has zero variance and then use the resulting estimate to obtain the optimal weights.

2.5 Methods for Estimating the Autocovariances of the Latent Process

Zeger (1988) suggested method of moment estimators for the autocorrelations of the latent process $\{\epsilon_t\}$. These are as follows:

$$\hat{\rho}_\epsilon(\tau) = \hat{\gamma}_\epsilon(\tau)/\hat{\sigma}^2_Z$$

with autocovariance estimates given by

$$\hat{\gamma}_\epsilon(\tau) = \sum_{t=\tau+1}^{n} \{(Y_t - \hat{\mu}_t)(Y_{t-\tau} - \hat{\mu}_{t-\tau})\} / \sum_{t=\tau+1}^{n} \hat{\mu}_t \hat{\mu}_{t-\tau}.$$

The estimates of variance and autocovariances are not guaranteed to form a non-negative definite sequence and therefore the autocorrelations are not guaranteed to be less than one in absolute values.

Brannas and Johansson (1994) suggest OLS type estimators derived by letting $\tilde{Y}_t = (Y_t - \hat{\mu}_t)$ and noting that from the autocovariance function of Y_t in Section 2.1,

$$E(\tilde{Y}_t \tilde{Y}_{t+\tau}) \approx \rho_\epsilon(\tau)\sigma_\epsilon^2 \mu_t \mu_{t+\tau},$$

so that regressing $\tilde{Y}_t \tilde{Y}_{t+\tau}$ on $\mu_t \mu_{t+\tau}$ leads to estimates

$$\hat{\gamma}_{OLS,\epsilon}(\tau) = \sum_{t=\tau+1}^{n} \hat{\mu}_t \hat{\mu}_{t+\tau} \{(Y_t - \hat{\mu}_t)(Y_{t-\tau} - \hat{\mu}_{t-\tau})\} / [\hat{\sigma}^2_{OLS} \sum_{t=\tau+1}^{n} \hat{\mu}_t^2 \hat{\mu}_{t-\tau}^2],$$

where $\hat{\sigma}^2_{OLS}$ is as defined in Section 2.4.

In both the Zeger definition and the least squares definition of Brannas and Johansson there is no attempt to standardize the individual terms in the summations for unequal variances. By analogy with the use of weights in forming estimates of the variance of the latent process some form of weighting could be useful in forming autocovariance estimates. We consider weighted estimates of autocovariances which are required to be unbiased for any underlying stationary latent process. Calculation of the variance of any such estimates will require estimation and knowledge of the autocovariances. However, in the case where the latent process is a sequence of independent random variables the variance of these weighted estimates is readily computable as we will show. Since the hypothesis of independence is of primary interest, obtaining minimum variance estimates is desirable.

Consider the following weighted estimates:

$$\hat{\gamma}_{\epsilon,W}(h) = \frac{1}{\sum_{t=1}^{n-h} W_t W_{t+h}} \sum_{t=1}^{n-h} W_t W_{t+h} \hat{V}_t \hat{V}_{t+h},$$

where $\hat{V}_t = \tilde{e}_t / \sqrt{\hat{\mu}_t}$. These estimates are approximately unbiased for the true covariance since the individual terms satisfy

$$\begin{aligned} E(\hat{V}_t \hat{V}_{t+h}) &\approx (\mu_t \mu_{t+h})^{-1} \mathrm{Cov}(Y_t, Y_{t+h}) \\ &= \gamma_\epsilon(h). \end{aligned}$$

Because of this unbiasedness, use of V_t as the basis for constructing weighted estimates seems reasonable. The approximate variance of these estimates under the assumption that the latent process is white noise is:

$$\mathrm{Var}(\hat{\gamma}_{\epsilon,W}(h)) \approx \frac{1}{(\sum_{t=1}^{n-h} W_t W_{t+h})^2} \sum_{t=1}^{n-h} W_t^2 W_{t+h}^2 (\sigma_\epsilon^2 + 1/\mu_t)(\sigma_\epsilon^2 + 1/\mu_{t+h}).$$

Alternatively the following weights can be derived directly from the finite sample approximation as

$$W_t^* = (\hat{\sigma}_\epsilon^2 + 1/\hat{\mu}_t)^{-1}.$$

Zeger's estimates use weights $W_t^Z = \hat{\mu}_t$ which leads to an approximate variance

$$\mathrm{Var}(\hat{\gamma}_{\epsilon,W^Z}(h)) \approx \frac{1}{(\sum_{t=1}^{n-h} \hat{\mu}_t \hat{\mu}_{t+h})^2} \sum_{t=1}^{n-h} \hat{\mu}_t^2 \hat{\mu}_{t+h}^2 (\hat{\sigma}_\epsilon^2 + 1/\hat{\mu}_t)(\hat{\sigma}_\epsilon^2 + 1/\hat{\mu}_{t+h}).$$

A similar expression can be derived for the Brannas-Johansson estimates in which the weights are $W_t^{BJ} = \hat{\mu}_t^2$.

Whatever the weighting scheme used the standard errors of the estimated covariances can be estimated, under the assumption that the latent process is an independent sequence, by

$$s.e.(\hat{\gamma}_{\epsilon,W}(h)) \approx \left[\frac{1}{(\sum_{t=1}^{n-h} \hat{W}_t \hat{W}_{t+h})^2} \sum_{t=1}^{n-h} \hat{W}_t^2 \hat{W}_{t+h}^2 (\hat{\sigma}_\epsilon^2 + 1/\hat{\mu}_t)(\hat{\sigma}_\epsilon^2 + 1/\hat{\mu}_{t+h}) \right]^{1/2}.$$

Integral approximations and simulations reported in Davis, Dunsmuir, and Wang (1997) lead to the following observations.

REMARK 2.4 The optimal estimates outperform the Zeger estimates which in turn outperform the BJ estimates. The performance advantages decrease as the degree of autocorrelation increases positively.

REMARK 2.5 When there is substantial autocorrelation present there is also substantial bias in the estimates of the $\gamma_\epsilon(h)$ — this occurs in all estimates to a similar degree. This would impact the estimation of the correct asymptotic variance in the GLM estimates of $\boldsymbol{\beta}$ of the above theorem and would tend to lead to underestimating the correct standard errors of the GLM estimates when there is substantial positive serial correlation present.

REMARK 2.6 However for the estimation of $\hat{\rho}_\epsilon(1)$ the bias is not as severe. For some purposes such as construction of a suitable correlation model for use in an efficient estimation procedure this reduction in bias is a good property. However all biases depend on the form of the regressor with worse bias for the linear trend regression than the cosine regression. This indicates that any bias adjustment procedure should account for the form of the regression function—see later for some proposals.

REMARK 2.7 The asymptotic formulae for $s.e.(\hat{\gamma}_{\epsilon,W}(h))$ give unbiased estimates of the simulated standard deviations when there is no serial dependence, i.e., in the situation that they are derived under. This means that this formula will be useful in calculating an overall test of autocovariance (see Section 2.6 below). Further the asymptotic formula provide reasonable formulae for estimating the null hypotheses standard deviations even in these cases where the null hypothesis is not true.

2.6 Tests for Zero Autocovariance in the Latent Process

In Brannas and Johansson (1994) the performance of the standard Box-Pierce and Ljung-Box portmanteau statistics is investigated. Three types of residuals (Pearson, Anscombe and their own proposal) are used to define the autocorrelations in the standard fashion. However there will be some inaccuracies in any such procedures unless the standard errors are properly calculated. The problem with the correlated Poisson model is that the variance and covariances have different forms of dependence on the mean function μ_t and there is no single normalization of residuals which will simultaneously eliminate this dependence from the variance and from the covariance terms required to construct autocorrelations. Hence the usual normalisation in the Box-Pierce and Ljung-Box portmanteau statistics will be incorrect.

Brannas and Johansson (1994) observe that in the portmanteau statistics constructed in this fashion "the sizes are significantly too high". One explanation of their results might be that the Box-Pierce and Ljung-Box statistics need to be adjusted to account for the correct standard errors. To illustrate this consider the use of Pearson residuals in the standard formulae for autocorrelation estimates which are, because the sample mean of the $\tilde{\epsilon}_t$ is zero, equal to

$$\hat{\rho}_{\epsilon,P}(h) = \frac{\sum_{t=1}^{n-h} \tilde{\epsilon}_t \tilde{\epsilon}_{t+h}}{\sum_{t=1}^{n} \tilde{\epsilon}_t^2}.$$

Using the types of formulae that are given above it can be shown that, under the assumption that the latent process is an iid sequence with mean 1 and variance σ_ϵ^2,

$$\mathrm{Var}(\hat{\rho}_{\epsilon,P}(h)) \approx [\frac{1}{n}\sum_{t=1}^{n}(1+\mu_t\sigma_\epsilon^2)]^{-2}[\frac{1}{n^2}\sum_{t=1}^{n-h}(1+\mu_t\sigma_\epsilon^2)(1+\mu_{t+h}\sigma_\epsilon^2)] =: V_h/n.$$

These can be readily estimated using the fitted mean and estimated variance. Also, for understanding the theoretical aspects of these estimated autocorrelations the true values of the mean function μ_t and variance σ_ϵ^2 can be used. For the cosine regression introduced above the value of $V_h/n = 1.19/n$ for $h = 1$ compared with the Ljung-Box value of $(n-h)/(n(n+2)) = 0.97/n$ and $V_h/n = 0.72/n$ for $h = 6$ compared with the Ljung-Box value of $(n-h)/(n(n+2)) = 0.92/n$. While these individual differences are substantial their effect will not be so large in a test statistic constructed using a number of autocorrelation estimates. In particular for this cosine regression case the mean of the V_h/n over the first 13 lags is $0.97/n$ while that of $(n-h)/(n(n+2))$ is $0.91/n$ indicating very little average difference in the correct variances for the autocorrelations based in Pearson residuals and the incorrect variances implicit in the Ljung-Box statistic. Indeed, in this case, the use of Pearson residuals in standard estimates of autocorrelations leads to a Ljung-Box statistic with good nominal coverage. In other situations the average of the first several values of $n\mathrm{Var}(\hat{\rho}_{\epsilon,P}(h))$ may not be so close to unity and there will be a need to properly calculate the asymptotic variances. However, based on the results presented above this is not difficult to do computationally, and, because it is always the correct procedure, we would recommend its universal use in these models.

Although these effects are small there will be circumstances where they could be large enough to badly distort the inference about serial autocorrelations using the classical (unadjusted) procedures from time series analysis of stationary sequences. To be accurate under a wide range of regression functions the classical statistical procedures for testing for the existence of autocorrelation should really to be modified in the present situation. Under the assumption that the latent process is white noise with positive variance the $\hat{\gamma}_{\epsilon,W}(h)$ are asymptotically distributed as independent normal random variables with standard deviations estimated by $s.e.(\hat{\gamma}_{\epsilon,W}(h))$ derived above. Hence a test statistic for zero autocovariance can be constructed based on the first several autocovariance estimates. The following statistic (analogous to the Box-Jenkins's portmanteau statistic) is proposed for testing for serial dependence in the mean of the observed count time series. For a given maximum lag length L define

$$H^2 = \sum_{h=1}^{L}[\hat{\gamma}_\epsilon(h)/s.e.(\hat{\gamma}_\epsilon(h))]^2.$$

Under the hypothesis of independence H^2 will have an approximate χ^2 distribution on L df. The following simulations indicate that this is a reasonable approximation.

Simulation results comparing the performance of H^2 using the estimates of Zeger, Brannas and Johansson and the optimally weighted estimates (i.e. those using W_t^*) are presented in detail in Davis, Dunsmuir, and Wang (1997) from which the following partial results are obtained.

Case 1: Type I Errors. Linear Trend, $\phi = 0, \gamma = 0.6931, \sigma_\epsilon^2 = 1.00$.

In this case a latent process is present but there is no autocovariance in it. The following results are based on 1000 replications in the simulations and the first 10 autocovariances.

	Observed	Observed	Power against		
$\alpha = 0.05$	Type I error	95%-ile	$\phi = 0.2$	$\phi = 0.4$	$\phi = 0.6$
$P(H_{OPT}^2 > \chi_{10,(1-\alpha)}^2)$	.038	17.3	.099	.330	.736
$P(H_Z^2 > \chi_{10,(1-\alpha)}^2)$	.031	16.9	.089	.280	.647
$P(H_{BJ}^2 > \chi_{10,(1-\alpha)}^2)$	.025	14.5	.118	.267	.590
Ljung-Box using e_t	.052	18.4	.093	.321	.714

Case 2: Type I Errors. Cosine Trend $\phi = 0, \gamma = 0.6931, \sigma_\epsilon^2 = 1.00$.

In this case a latent process is present but there is no autocovariance in it. The following results are based on 1000 replications in the simulations and the first 10 autocovariances.

	Observed	Observed	Power against		
$\alpha = 0.05$	Type I error	95%-ile	$\phi = 0.2$	$\phi = 0.4$	$\phi = 0.6$
$P(H_{OPT}^2 > \chi_{10,(1-\alpha)}^2)$	.045	18.0	.091	.286	.844
$P(H_Z^2 > \chi_{10,(1-\alpha)}^2)$	.061	19.4	.070	.177	.806
$P(H_{BJ}^2 > \chi_{10,(1-\alpha)}^2)$	.128	25.0	.052	.108	.645
Ljung-Box using e_t	.044	17.8	.089	.293	.854

The "optimal" estimates used in the test statistic provide approximately similar performance to the Ljung-Box statistic. Both of these are superior to the use of Zeger's estimates which in turn is superior to the BJ estimates. Since H_{OPT}^2 is based on the correct variances of the estimates autocovariances while the Ljung-Box statistic based on Pearson residuals does not we would recommend the use of H_{OPT}^2 universally.

2.7 Bias Adjustment for Estimation of Autocovariances

Note that the use of Pearson residuals for estimation of autocorrelations in the non-null case (i.e., where there is serial correlation) is biased in general because

$$E(\frac{1}{n}\sum_{t=1}^{n-h}\tilde{\epsilon}_t\tilde{\epsilon}_{t+h}) \approx [\frac{1}{n}\sum_{t=1}^{n-h}\mu_t^{1/2}\mu_{t+h}^{1/2}]\gamma_\epsilon(h)$$

while

$$\begin{aligned} E(\frac{1}{n}\sum_{t=1}^{n}\tilde{\epsilon}_t^2) &\approx \frac{1}{n}\sum_{t=1}^{n}\mu_t^{-1}[\mu_t + \sigma_\epsilon^2\mu_t^2] \\ &= 1 + \left(\frac{1}{n}\sum_{t=1}^{n}\mu_t\right)\sigma_\epsilon^2. \end{aligned}$$

In the case when $\mu_t \equiv \mu$, we have $E(\frac{1}{n}\sum_{t=1}^{n-h}\tilde{\epsilon}_t\tilde{\epsilon}_{t+h}) \approx \mu\gamma_\epsilon(h)$ and $E(\frac{1}{n}\sum_{t=1}^{n}\tilde{\epsilon}_t^2) \approx (1+\mu\sigma_\epsilon^2)$ giving

$$\hat{\rho}_{\epsilon,P}(h) \approx \frac{\mu\sigma_\epsilon^2}{(1+\mu\sigma_\epsilon^2)}\rho_\epsilon(h),$$

which is as observed for the autocorrelation of the Y_t process in Section 2.1 above. For the cosine regression model used above in the bias factors for autocorrelation estimation are 0.82 at $h = 1$ and 0.58 at $h = 6$.

While this theory suggest we can expect bias in the standard ACF estimates using Pearson residuals the theoretically asymptotically unbiased estimates $\hat{\gamma}_Z(h)$ and $\hat{\gamma}_{OPT}(h)$ are in fact also severely biased if the true autocorrelations are non-zero. Consider $\phi = 0.9$ and the cosine regression model. Simulations reported in Davis, Dunsmuir, and Wang (1997) compare the bias properties of the autocovariances and the autocorrelations estimated using optimal weighting and using the standard formulae based on Pearson residuals. The following simulation results are based on 1000 replicates of a series of length 100.

Lag	Autocovariances			Autocorrelations		
h	True $\gamma_\epsilon(h)$	$\hat{\gamma}_{\epsilon,W^*}(h)$	$\hat{\gamma}_{\epsilon,P}(h)$	True $\rho_\epsilon(h)$	$\hat{\rho}_{\epsilon,W^*}(h)$	$\hat{\rho}_{\epsilon,P}(h)$
0	1.0	.60	3.18	1.0	1.0	1.0
1	.87	.50	1.77	0.87	.83	.49
6	.45	.21	0.59	0.45	.33	.15

Clearly use of Pearson residuals in the standard autocovariance and autocorrelation methods ($\hat{\gamma}_{\epsilon,P}(h)$ and $\hat{\rho}_{\epsilon,P}(h)$) are badly biased and much more so than the optimal weighted estimates ($\hat{\gamma}_{\epsilon,W^*}(h)$ and $\hat{\rho}_{\epsilon,W^*}(h)$). Use of "hat" adjusted Pearson residuals scarcely altered these results. Likewise the use of Zeger estimates gives almost identical bias in both the autocovariance and the autocorrelation estimates as in the optimal estimates. Note that the bias in the Pearson residual based autocorrelation estimates is even greater than predicted by the above theoretical calculation.

Estimates of autocovariances using Zeger's method or the optimally weighted method are substantially biased towards zero when the true autocovariances were positive as arises for example in the autoregression with positive parameter. For the purposes of good estimates of asymptotic covariance matrix for the GLM estimates of the regression parameters removal of bias from the estimates of variance and autocovariance is warranted.

Standard approaches to adjusting for bias include use of $n - p$ or use of the 'hats' h_t to adjust for the use of fitted means in place of true means when calculating residuals. We attempted this, but similarly to their use in the Pearson residual based a.c.f. they had little effect on the bias. Also different types of residuals, as suggested by Brannas and Johansson could be used. But we doubt this will overcome the problem.

The Zeger and optimally weighted estimates would be unbiased if the true mean μ_t was used in their definition. However as we will demonstrate the effect of variability in the estimates of $\boldsymbol{\beta}$ can have a substantial impact on bias. The traditional adjustment procedures do not allow for the impact of substantial positive autocorrelation in the latent process in the standard errors of regression estimates which in turn can effect bias because the estimates mean for these Poisson models is an exponential function of a linear function in the estimated $\boldsymbol{\beta}$. Accordingly it is the *variance* of $\hat{\boldsymbol{\beta}}$ that is important to the *bias* in $\hat{\mu}$. Fortunately, simple expansions to adjust for use of $\hat{\mu}_t$ in place of μ_t in the above definitions can be derived in a relatively straightforward way using the asymptotic distribution for $\hat{\boldsymbol{\beta}}$ derived above.

Bias adjusted "optimally" weighted estimates, proposed in Davis, Dunsmuir, and Wang (1997), are defined as follows:

$$\hat{\sigma}^2_{\epsilon,UB} = \frac{\sum_{t=1}^n (W_t^*/\hat{\mu}_t)^2[(Y_t - \hat{\mu}_t)^2 + \hat{\mu}_t^2 \mathbf{x}_t^T \hat{\Omega} \mathbf{x}_t - \hat{\mu}_t]}{\sum_{t=1}^n (W_t^*/\hat{\mu}_t)^2 \hat{\mu}_t^2 \exp(-2\mathbf{x}_t^T \hat{\Omega} \mathbf{x}_t)},$$

where

$$W_t^{*2} = 1/Var(E_t) = \mu_t^4/B_t,$$

and

$$\hat{\gamma}_{\epsilon,UB}(h) = \frac{\sum_{t=1}^{n-h}(W_t^*/\hat{\mu}_t)(W_{t+h}^*/\hat{\mu}_{t+h})\{(Y_t-\hat{\mu}_t)(Y_{t+h}-\hat{\mu}_{t+h}) + \hat{\mu}_t \mathbf{x}_t^T \hat{\Omega} \mathbf{x}_{t+h}\hat{\mu}_{t+h}\}}{\sum_{t=1}^{n-h}(W_t^*/\hat{\mu}_t)(W_{t+h}^*/\hat{\mu}_{t+h})\hat{\mu}_t\hat{\mu}_{t+h}\exp\{-(\mathbf{x}_t+\mathbf{x}_{t+h})^T\hat{\Omega}(\mathbf{x}_t+\mathbf{x}_{t+h})/2\}},$$

where

$$W_t^* = (\hat{\sigma}_\epsilon^2 + 1/\hat{\mu}_t)^{-1}.$$

Case 1: $\phi = 0.9$, linear regression.

The following tables contain simulation results for the autocovariances.

		Mean		S.D.		RMSE	
Lag	True	$\hat{\gamma}_{\epsilon,W^*}(h)$	$\hat{\gamma}_{\epsilon,UB}(h)$	$\hat{\gamma}_{\epsilon,W^*}(h)$	$\hat{\gamma}_{\epsilon,UB}(h)$	$\hat{\gamma}_{\epsilon,W^*}(h)$	$\hat{\gamma}_{\epsilon,UB}(h)$
0	1.00	.48	.68	.28	.52	.59	.61
1	0.87	.38	.56	.25	.47	.54	.56
6	0.45	.10	.21	.14	.26	.38	.35

For the autocorrelation estimates, the simulation results reported below are conditional on $S_a > 1.645$ so that the null hypothesis of no latent process is rejected at the 5% level. This was done to eliminate values of autocorrelations badly effected by zero or near zero variance estimates. In 2 out of 1000 replications S_a was less than 1.645.

		Mean		S.D.		RMSE	
Lag	True	$\hat{\rho}_{\epsilon,W^*}(h)$	$\hat{\rho}_{\epsilon,UB}(h)$	$\hat{\rho}_{\epsilon,W^*}(h)$	$\hat{\rho}_{\epsilon,UB}(h)$	$\hat{\rho}_{\epsilon,W^*}(h)$	$\hat{\rho}_{\epsilon,UB}(h)$
1	0.87	.78	.80	.18	.16	.20	.17
6	0.45	.16	.27	.25	.23	.38	.29

Case 2: $\phi = 0.9$ cosine regression. 1000 replications.

The following tables contain simulation results for the autocovariances and the autocorrelations.

		Mean		S.D.		RMSE	
Lag	True	$\hat{\gamma}_{\epsilon,W^*}(h)$	$\hat{\gamma}_{\epsilon,UB}(h)$	$\hat{\gamma}_{\epsilon,W^*}(h)$	$\hat{\gamma}_{\epsilon,UB}(h)$	$\hat{\gamma}_{\epsilon,W^*}(h)$	$\hat{\gamma}_{\epsilon,UB}(h)$
0	1.00	.61	.79	.42	.64	.57	.67
1	0.87	.51	.67	.37	.56	.52	.60
6	0.45	.21	.29	.23	.35	.33	.38

Results for autocorrelation estimates. These results are conditional on $S_a > 1.645$ (in 9 out of 1000 replications S_a was less than 1.645).

		Mean		S.D.		RMSE	
Lag	True	$\hat{\rho}_{\epsilon,W^*}(h)$	$\hat{\rho}_{\epsilon,UB}(h)$	$\hat{\rho}_{\epsilon,W^*}(h)$	$\hat{\rho}_{\epsilon,UB}(h)$	$\hat{\rho}_{\epsilon,W^*}(h)$	$\hat{\rho}_{\epsilon,UB}(h)$
1	0.87	0.83	0.85	.18	.18	.19	.18
6	0.45	0.32	0.33	.31	.28	.33	.30

Note that the bias improved versions of the estimates of the autocovariances do indeed have better bias properties but at the expense of higher variance so that in RMSE terms they perform worse than the original unadjusted estimates. However the bias is still unacceptably high and further improvements should be sought.

From the above tables we also see that the estimates of autocorrelations have better bias properties. This means that for model identification and estimation purposes for which autocorrelations only are required the bias adjusted estimates or the nonadjusted estimates will be adequate. Note that the bias adjusted estimates of autocorrelations are slightly better in S.D. and RMSE terms than the unadjusted estimates. However for purposes in which an unbiased estimate of scale is required even the bias adjusted estimates of $\hat{\sigma}_\epsilon^2$ is biased towards 0. This will impact the estimation of standard errors of $\hat{\beta}_{GLM}$ and the method of Zeger for example.

2.8 Estimation

Zeger's treatment of the estimation of the latent process model is based on a quasi-likelihood approach used to correct for serial correlation in the latent process $\{\epsilon_t\}$. Assumptions on the distributional properties of this process are not explicitly stated but for much of his treatment these are not required. However for the alternative specification in terms of the $\{\delta_t\}$ process the requirement that these be normally distributed is made quite explicitly in the treatment given in Chan and Ledolter (1995). Indeed as we will demonstrate later, it is difficult to obtain the exact and large sample statistical properties required for inference in this log-normal specification without the assumption of normality.

Detection of autocorrelation using the observed count process, Y_t, is not straightforward as we showed in Section 2.1. Typically use of the autocorrelations for Y_t will lead to underestimates of the true degree of autocorrelation in the latent process ϵ_t. This is noted in Zeger (1988). Also, in practice, the regression on $\mathbf{x}_t$ will need to be performed before attempting to estimate this autocorrelation. However such estimation should adjust for autocorrelation in order to arrive at correct inferences. Estimation of the parameters $\boldsymbol{\theta}$ in the model by direct numerical maximization of the likelihood function is difficult since the likelihood cannot be written down in closed form. (For model (2.1), from (1.3) the likelihood is the n-fold integral,

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\{\sum_{t=1}^{n}(\mathbf{x}_t^T\boldsymbol{\beta}y_t-\epsilon_t e^{\mathbf{x}_t^T\boldsymbol{\beta}})\}(\prod_{t=1}^{n}\epsilon_t^{y_t})L(\boldsymbol{\theta};\boldsymbol{\epsilon}^{(n)})\,(d\epsilon_1\cdots d\epsilon_n)/\prod_{i=1}^{n}(y_i!),$$

where $L(\boldsymbol{\theta};\boldsymbol{\epsilon}^{(n)})$ is the likelihood based on $\epsilon_1,\ldots,\epsilon_n$.) To overcome this difficulty, Chan and Ledolter (1995) proposed an algorithm, called Monte Carlo EM (MCEM), whose iterates $\boldsymbol{\theta}^{(i)}$ converge to the maximum likelihood estimate. To apply this algorithm, first note that the conditional distribution of $\mathbf{Y}^{(n)}$ given $\boldsymbol{\epsilon}^{(n)}$ does not depend on $\boldsymbol{\theta}$ so that the likelihood based on the complete data $(\boldsymbol{\epsilon}^T,\mathbf{Y}^T)$ is given by

$$L(\boldsymbol{\theta};\boldsymbol{\epsilon}^{(n)},\mathbf{Y}^{(n)})=f(\mathbf{Y}^{(n)}|\boldsymbol{\epsilon}^{(n)})L(\boldsymbol{\theta};\boldsymbol{\epsilon}^{(n)}).$$

The E-step of the algorithm requires calculation of

$$\begin{aligned}Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i)}) &= E_{\boldsymbol{\theta}^{(i)}}\left(\ln L(\boldsymbol{\theta};\boldsymbol{\epsilon}^{(n)},\mathbf{Y}^{(n)})|\mathbf{Y}\right)\\ &= E_{\boldsymbol{\theta}^{(i)}}\left(\ln f(\mathbf{Y}^{(n)}|\boldsymbol{\epsilon}^{(n)})|\mathbf{Y}\right)+E_{\boldsymbol{\theta}^{(i)}}\left(\ln L(\boldsymbol{\theta};\boldsymbol{\epsilon}^{(n)})|\mathbf{Y}\right).\end{aligned}$$

We delete the first term from the definition of Q, since it is independent of $\boldsymbol{\theta}$ and hence plays no role in the M-step of the EM algorithm. The new Q is redefined as

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i)})=E_{\boldsymbol{\theta}^{(i)}}\left(\ln L(\boldsymbol{\theta};\ \boldsymbol{\epsilon}^{(n)})|\mathbf{Y}\right). \tag{2.11}$$

Even with this simplification, direct calculation of Q is still intractable. Suppose for the moment that it is possible to generate replicates of $\epsilon^{(n)}$ from the conditional distribution of $\epsilon^{(n)}$ given $\mathbf{Y}^{(n)}$ when $\theta = \theta^{(i)}$. If we denote m separate replicates of $\epsilon^{(n)}$ by $\epsilon_1^{(n)}, \ldots, \epsilon_m^{(n)}$, then a Monte Carlo approximation to Q in (2.11) is given by

$$Q_m(\theta|\theta^{(i)}) = \frac{1}{m}\sum_{j=1}^{m} \ln L(\theta; \epsilon_j^{(n)}).$$

The M-step is easy to carry out using Q_m in place of Q (especially if we condition on $\epsilon_1 = 0$ in all the simulated replicates) since L is just the Gaussian likelihood of the regression model with AR(p) noise. The difficult steps in the algorithm are the generation of replicates of $\epsilon^{(n)}$ given $\mathbf{Y}^{(n)}$ and the choice of m. Chan and Ledolter (1995) discuss the use of the Gibb's sampler for generating the desired replicates and give some guidelines on the choice of m.

In their analyses of the polio data, Zeger (1988) and Chan and Ledolter (1995) included as regression components an intercept, a slope, and harmonics at periods of 6 and 12 months. Specifically, they took $\epsilon_t = \exp(\delta_t), \delta_t = \phi\delta_{t-1} + z_t, z_t$ is IID $N(0, \sigma^2)$, and

$$\mathbf{x}_t = (1, t/1000, \cos(2\pi t/12), \sin(2\pi t/12), \cos(2\pi t/6), \sin(2\pi t/6))^T.$$

The implementation of Chan and Ledolter's MCEM method by Kuk and Cheng (1994) gave estimates $\hat{\beta} = (.247, -3.871, .162, -.482, .414, -.011)^T$, $\hat{\phi} = .648$, and $\hat{\sigma}^2 = .281$. The negative coefficient of $t/1000$ indicates a slight downward trend in the monthly number of polio cases.

2.8.1 The Likelihood Function

Suppose $\{\epsilon_t\}$ is the latent process in a parameter-driven model

$$Y_t|\epsilon_t, \mathbf{x}_t \sim \mathcal{P}(\epsilon_t \exp\{\mathbf{x}_t^T\beta\}),$$

where $\epsilon_t = \exp(\delta_t)$ and $\{\delta_t\}$ is an autoregressive process of order p (AR(p)). That is $\{\delta_t\}$ satisfies the recursions

$$\delta_t = \phi_1\delta_{t-1} + \cdots + \phi_p\delta_{t-p} + z_t, \quad z_t \sim \text{IID}(0, \sigma^2). \tag{2.12}$$

The likelihood of the complete data $(\mathbf{y}, \delta) = (y_1, ..., y_n; \delta_1, ..., \delta_n)$ is given by

$$\begin{aligned} f(\mathbf{y}; \delta) &= f(\mathbf{y}|\delta) f(\delta) \\ &= \Pi_{i=1}^n f(y_i|\delta_i) f(\delta) \\ &= [\Pi_{i=1}^n \tfrac{\exp\{-\exp(\delta_i + \mathbf{x}_i^T\beta)\}\exp\{(\delta_i + \mathbf{x}_i^T\beta)y_i\}}{y_i!}] \tfrac{\exp\{-\frac{1}{2}\delta^T V\ \delta\}}{(2\pi)^{n/2}|V^{-1}|^{1/2}} \\ &= \tfrac{|V|^{1/2}}{C_1} \exp\{-\textstyle\sum_{i=1}^n \exp(\delta_i + \mathbf{x}_i^T\ \beta) + \sum_{i=1}^n (\delta_i + \mathbf{x}_i^T\ \beta) y_i - \frac{1}{2}\ \delta^T V \delta\} \\ &= \tfrac{|V|^{1/2}}{C_1} \exp\{-(\mathbf{e}^{\delta})^T \mathbf{e}^{X\ \beta} + \mathbf{y}^T\delta + \ \mathbf{y}^T X\beta - \tfrac{1}{2}\delta^T V\delta\} \end{aligned} \tag{2.13}$$

where $C_1 = (2\pi)^{n/2}(\Pi_{i=1}^n y_i!)$, $\mathbf{e}^{\delta} = (e^{\delta_1}, ..., e^{\delta_n})^T$ and $\mathbf{e}^{X\ \beta} = (e^{\mathbf{x}_1^T\beta}, ..., e^{\mathbf{x}_n^T\beta})^T$.

Our objective is to estimate the model parameters $\theta = (\beta^T, \phi^T, \sigma^2)^T$, where $\phi = (\phi_1, \ldots, \phi_p)^T$. As noted above, the likelihood of the observed data $\mathbf{y}$ does not have a simple closed form and maximum likelihood estimation is intractable. Instead, we approximate the likelihood of the complete data ($\mathbf{y}$, δ) by a distribution for which the δ are easily integrated in order to compute the marginal distribution of $\mathbf{y}$. We then estimate the model parameters by maximizing the approximate likelihood.

2.8.2 The Approximate Likelihood

A Taylor expansion at $\delta_0 = (\delta_1^{(0)}, \cdots, \delta_n^{(0)})$ on the term $(\mathbf{e}^{\delta})^T \mathbf{e}^{X\beta}$ in equation (2.13) gives

$$(\mathbf{e}^{\delta})^T \mathbf{e}^{X\beta} = \mathbf{b}_0{}^T \mathbf{e}^{X\beta} + (\delta - \delta_0)^T K \mathbf{b}_0 + \frac{1}{2}(\delta - \delta_0)^T BK(\delta - \delta_0) \tag{2.14}$$

where $\mathbf{b}_0 = (e^{\delta_1^{(0)}}, ..., e^{\delta_n^{(0)}})^T$, $B = \text{diag}(e^{\delta_1^{(0)}}, ..., e^{\delta_n^{(0)}})$, $K = \text{diag}(e^{\mathbf{x}_1^T\beta}, ..., e^{\mathbf{x}_n^T\beta})$. Let

$$\tilde{\mathbf{y}} = \mathbf{y} - K\mathbf{b}_0 + BK\,\delta_0, \tag{2.15}$$

the approximate likelihood of the complete data $(\mathbf{y}, \delta)$ is then given by

$$\begin{aligned}
f_a(\mathbf{y}, \delta) &= \frac{|V|^{1/2}}{C_1} \exp\{\mathbf{y}^T\delta + \mathbf{y}^T X\,\beta - \tfrac{1}{2}\delta^T V\,\delta \\
&\quad -[\mathbf{b}_0{}^T \mathbf{e}^{X\,\beta} + (\delta - \delta_0)^T K\mathbf{b}_0 + \tfrac{1}{2}(\delta - \delta_0)^T BK(\delta - \delta_0)]\} \\
&= \frac{|V|^{1/2}}{C_1} \exp\{(\mathbf{y} - K\mathbf{b}_0 + BK\,\delta_0)^T\delta - \tfrac{1}{2}\delta^T(BK + V)\delta \\
&\quad -\tfrac{1}{2}\delta_0{}^T BK\delta_0 + \delta_0{}^T K\mathbf{b}_0 + \mathbf{y}^T X\beta - \mathbf{b}_0^T \mathbf{e}^{X\,\beta}\} \\
&= \frac{|V|^{1/2}}{C_1} \exp\{\tilde{\mathbf{y}}^T\delta - \tfrac{1}{2}\delta^T(BK + V)\delta \\
&\quad -\tfrac{1}{2}\delta_0{}^T BK\delta_0 + \delta_0{}^T K\mathbf{b}_0 + \mathbf{y}^T X\beta - \mathbf{b}_0^T \mathbf{e}^{X\,\beta}\} \\
&= \frac{|V|^{1/2}}{C_1} \exp\{-\tfrac{1}{2}\left[\delta - (BK + V)^{-1}\tilde{\mathbf{y}}\right]^T (BK + V)\left[\delta - (BK + V)^{-1}\tilde{\mathbf{y}}\right] \\
&\quad +\tfrac{1}{2}\tilde{\mathbf{y}}^T(BK + V)^{-1}\tilde{\mathbf{y}} - \tfrac{1}{2}\,\delta_0{}^T BK\delta_0 + \delta_0{}^T K\mathbf{b}_0 + \mathbf{y}^T X\,\beta - \mathbf{b}_0\mathbf{e}^{X\beta}\}.
\end{aligned}$$

So the conditional distribution of δ given $\mathbf{y}$ is

$$\delta|\mathbf{y} \sim N\left((BK + V)^{-1}\tilde{\mathbf{y}}, (BK + V)^{-1}\right), \tag{2.16}$$

and the approximate distribution of $\mathbf{y}$ is

$$\begin{aligned}
f_a(\mathbf{y}) &= \frac{|V|^{1/2}}{|BK+V|^{1/2}(\Pi_{i=1}^n y_i!)} \exp\{\tfrac{1}{2}\tilde{\mathbf{y}}^T(BK + V)^{-1}\tilde{\mathbf{y}} \\
&\quad -\tfrac{1}{2}\delta_0{}^T BK\,\delta_0 + \delta_0{}^T K\mathbf{b}_0 + \mathbf{y}^T X\beta - \mathbf{b}_0^T \mathbf{e}^{X\beta}\}.
\end{aligned} \tag{2.17}$$

The matrix V^{-1} is the covariance matrix of the observation vector $(\delta_1, \cdots, \delta_n)$ from the process $\{\delta_t\}$. In order to get the approximate distribution of $\mathbf{y}$, we need to calculate V, $|V|$, $|BK + V|$, and $(BK + V)^{-1}\tilde{\mathbf{y}}$.

The calculation of $|BK+V|$ and $(BK+V)^{-1}\tilde{\mathbf{y}}$ are based on the innovations algorithm (Brockwell and Davis (1991)) that avoids inverting the matrix $(BK + V)$ directly. We treat $(BK + V)$ as a covariance matrix of $\tilde{\mathbf{y}}$ and decompose it as

$$BK + V = CDC^T,$$

where C is a lower triangular matrix with all the diagonal elements 1, $D = \text{diag}(v_0, v_1, \cdots, v_{n-1})$, and v_i is the mean squared error of the one-step predictor of $\tilde{y}_i$. Then the determinant of $(BK+V)$ is given by

$$|BK + V| = |C||D||C^T| = \Pi_{i=0}^{n-1} v_i.$$

Moreover the innovations algorithm can easily be adapted to compute the conditional mean, $E(\delta|\mathbf{y}) = (BK + V)^{-1}$ specified in (2.16).

The parameters of the model are then estimated as follows.

1. Fix initial values of $\delta = \delta^{(0)}$, $\phi = \phi^{(0)}$, and $\sigma^2 = \sigma^{2(0)}$;

2. For fixed $\delta^{(j)}, \phi^{(j)}$ and $\sigma^{2(j)}$, maximize $\mathbf{y}^T X\beta - \mathbf{b}_0^T e^{X\beta}$ to get $\beta^{(j+1)}$; this is comparable is Poisson regression.

3. For fixed $\delta^{(j)}$ and $\beta^{(j+1)}$, maximize $\log \frac{|V|}{|BK+V|} + \tilde{\mathbf{y}}^T (BK+V)^{-1}\tilde{\mathbf{y}}$ to find $\phi^{(j+1)}$ and $\sigma^{2(j+1)}$, the estimates of ϕ and σ^2 respectively;

4. For fixed $\beta^{(j+1)}, \phi^{(j+1)}$ and $\sigma^{2(j+1)}$, use $\delta^{(j+1)} = (BK+V)^{-1}\tilde{\mathbf{y}}$ iteratively to get the estimates of δ. After convergence, set $\delta^{(\infty)} = \delta^{(j+1)}$;

5. Increment j, go to step 2 and continue to convergence.

We applied this algorithm to the polio data. Using the same regression terms as before (see Section 2.2.2) i.e.,

$$\mathbf{x}_t = (1, t/1000, \cos(2\pi t/12), \sin(2\pi t/12), \cos(2\pi t/6), \sin(2\pi t/6))^T.$$

and an AR(1) model for the latent process $\{\delta_t\}$ we obtain

$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\phi}$	$\hat{\sigma}^2$
0.407	-4.236	0.153	-0.466	0.402	-0.008	0.664	0.244

These values are in reasonably close agreement with those cited earlier by Kuk and Cheng (1994) and require a great deal less computation. Note that these estimates are not directly comparable with the earlier analyses because the regression terms are not centered at January, 1976 in order to match the analysis of Kuk and Cheng.

Durbin and Koopman (1997) consider a Gaussian state equation and a non-Gaussian observation process. They approximate the density $p(y_t|\epsilon_t)$ by a Gaussian conditional density $g(y_t|\epsilon_t)$ to arrive at a Gaussian approximation to the likelihood. Simulation techniques are then used to compute the adjustment from the approximate likelihood to the correct likelihood. The conditional mode and expected value of ϵ_t are also calculated efficiently using their method. An interesting future research question is to compare the performance of their methods with that proposed here and with the use of full simulation of the exact likelihood based on the Markov chain Monte Carlo ideas described above.

3 Observation Driven Models

3.1 Review of Existing Models

In this section we review various observation driven models for count data. In addition we offer some alternatives that have been considered in the literature to date.

In an early contribution, Harvey and Fernandes (1989) considered observations Y_t which have a Poisson distribution conditional upon a mean process ϵ_t. They specify the conditional density of ϵ_t conditional upon information up to time $t-1$ as that of a gamma distribution with parameters $a_{t|t-1}$ and $b_{t|t-1}$. This density corresponds to a conjugate prior for the Poisson distribution. The posterior density of ϵ_t given observations up to time t is also gamma with parameters a_t and b_t. The two sets of parameters are assumed to be link through the relations $a_{t|t-1} = \omega a_t$ and $b_{t|t-1} = \omega b_t$ for some $0 < \omega \leq 1$. Using this formulation the stochastic mechanism for the transition from ϵ_{t-1} to ϵ_t is defined implicitly. The log-likelihood for ω is defined and the forecast function $E(Y_{T+1}|\mathbf{Y}^{(\mathbf{T})})$ is shown to follow an exponentially weighted moving average of past Y_t's. Similar models, based on appropriate conjugate distributions, are developed for the binomial, multinomial and negative exponential observation distributions. Harvey and Fernandes also extend their methods to incorporate estimation of explanatory variables by modifying, in the Poisson case, the mean function to $\epsilon_t \exp(\mathbf{x}_t^T\beta)$ and apply this model to various time series of counts.

In a different direction, and one that we pursue in more detail, Zeger and Qaqish (1988) considered the following model for Poisson counts. Let

$$\mathbf{H}_t = (\mathbf{Y}^{(t-1)}, \mathbf{X}^{(t)})$$

be the past of the observed count process and the past and present of the regressor variables. Zeger and Qaqish (1988) assume that the conditional distribution of $Y_t|\mathbf{H}_t$ is Poisson with mean given by

$$\mu_t = \exp(\mathbf{x}_t^T\boldsymbol{\beta}) \prod_{i=1}^{p} \left[\frac{\max(Y_{t-i}, c)}{\exp(\mathbf{x}_{t-i}^T\boldsymbol{\beta})} \right]^{\gamma_i}, \tag{3.1}$$

where $c > 0$ is a constant which prevents $Y_{t-1} = 0$ becoming an absorbing state. Note that when $p = 1$, c can be interpreted as

$$c = P(Y_t > 0|Y_{t-1} = 0).$$

An alternative form considered by them is

$$\mu_t = \exp(\mathbf{x}_t^T\boldsymbol{\beta}) \prod_{i=1}^{p} \left[\frac{Y_{t-i} + c}{\exp(\mathbf{x}_{t-i}^T\boldsymbol{\beta}) + c} \right]^{\gamma_i},$$

in which, in the case $p = 1$, c is interpreted as an immigration rate adding to counts at every time point. In both these forms estimation of c could be problematic and the inferential theory for its estimation does not seem to have been worked out; because it represents an end point parameter this might be problematic. Thus the parameter c seems to be somewhat arbitrary and, apart from the interpretations just given for the case $p = 1$, it is not clear how to interpret it.

An alternative considered in Zeger and Qaqish (1988) is

$$\mu_t = \exp(\mathbf{x}_t^T\boldsymbol{\beta}) \exp(\sum_{i=1}^{p} \gamma_i Y_{t-i}). \tag{3.2}$$

Note that the Y_{t-i} enter without any form of mean correction or centering.

Zeger and Qaqish (1988) argue that model (3.1) is preferred on three criteria. In particular model (3.2) cannot be stationary unless, at least in the case $p = 1$, $\gamma_1 \leq 0$ thereby excluding the possibility of positive dependence.

Model (3.2) is applied to data in Fahrmeir and Tutz (1994). They offer some comments on the problems of stationarity (p. 195) but appear to ignore this issue when it comes to actually fitting an observation driven model to real data. When their form of the model is used with the asthma data the estimated long run mean and seasonal effects are badly distorted.

It might be thought that the difficulties with model (3.2) could be overcome by subtracting the deterministic mean from the Y_t to arrive at

$$\mu_t = \exp(\mathbf{x}_t^T\boldsymbol{\beta}) \exp(\sum_{i=1}^{p} \gamma_i (Y_{t-i} - \exp(\mathbf{x}_{t-i}^T\boldsymbol{\beta}))). \tag{3.3}$$

but in fact this will not lead to a stationary process as can be seen by using an argument similar to that used in Zeger and Qaqish.

3.2 Observation Driven Models in Standardized Errors

A different form of mean correction is proposed here which could result in the generation of stationary processes. However, as we will see, this is at the expense of a Markov (finite lag) structure for the Y_t process but not the conditional mean.

To introduce the model let, for $\lambda \geq 0$,

$$e_t = (Y_t - \mu_t)/\mu_t^{\lambda}$$

and

$$\log(\mu_t) = W_t = \mathbf{x}_t^T\boldsymbol{\beta} + \sum_{i=1}^{p} \gamma_i e_{t-i}$$

and assume that

$$Y_t|\mathbf{Y}^{(t-1)} \sim \mathcal{P}(\mu_t).$$

Note that the conditional mean, μ_t, is based on the whole past and so the process $\{Y_t\}$ is no longer Markov. However it could lead to stationary solutions although establishing this with rigor seems to be difficult. Note that $\{\mu_t\}$ is a pth order Markov chain. We discuss some of its properties below. However for the purposes of inference that discussion is not required.

Extensions to autoregressive moving average filters applied to past values of e_t can also be made as follows. Let

$$W_t = \mathbf{x}_t^T\boldsymbol{\beta} + \sum_{i=1}^{\infty} \tau_i e_{t-i},$$

where

$$\begin{aligned} \sum_{i=1}^{\infty} \tau_i z^i &= (1 - \sum_{i=1}^{q} \phi_i z^i)^{-1}(1 + \sum_{i=1}^{p} \gamma_i z^i) - 1 \\ &= \phi(z)^{-1}\gamma(z) - 1 \end{aligned}$$

and note that $\sum_{i=1}^{\infty} \tau_i e_{t-i}$ is the one step ahead predictor of the e_t based on an ARMA(q,p) model. In such a specification the infinite past is required but in practice this will not be available. Since the joint distribution of the e_t is not known initial conditions which conform to the distribution of values for $t \leq 0$ cannot be specified. The simplest proposal for practical applications might be to begin the recursions required for computing $\sum_{i=1}^{\infty} \tau_i e_{t-i}$ be setting $e_t = 0$ for $t \leq 0$.

Shephard (1995, unpublished) suggested a model quite similar to that proposed above. He presents an argument, based on a Taylor series linearization of the link function, for using $\lambda = 1$ in the definition of e_t at least for the Poisson case. We assume that τ_t is determined by information available at time t, i.e., τ_t depends on past observations $(y_{t-1}, y_{t-2}, \ldots, y_{t-p})$ and covariates up to time t. Define

$$\log(\mu_t) = \mathbf{x}_t^T\boldsymbol{\beta} + \sum_{i=1}^{q} \phi_i z_{t-i}$$

where

$$z_t = \mathbf{x}_t^T\boldsymbol{\beta} + \sum_{i=1}^{q} \phi_i z_{t-i} + e_t + \sum_{k=1}^{p} \gamma_k e_{t-k}.$$

This model is called the generalized linear ARMA (or GLARMA) model by Shephard (1995). Note that

$$z_t = \phi(B)^{-1}\mathbf{x}_t^T\boldsymbol{\beta} + \phi(B)^{-1}\gamma(B)e_t,$$

so that

$$\begin{aligned}\log(\mu_t) &= \mathbf{x}_t^T\boldsymbol{\beta} + \sum_{i=1}^{q}\phi_i z_{t-i} \\ &= \mathbf{x}_t^T\boldsymbol{\beta} + [1-\phi(B)][\phi(B)^{-1}\mathbf{x}_t^T\boldsymbol{\beta} + \phi(B)^{-1}\gamma(B)e_t] \\ &= \phi(B)^{-1}\mathbf{x}_t^T\boldsymbol{\beta} + [\phi(B)^{-1}-1]\gamma(B)e_t.\end{aligned}$$

This formulation differs from that proposed by us above in the following ways:

1. It uses $\lambda = 1$. This means that the e_t do not have unit variance but instead have variance that depends on the mean function μ_t. For $\lambda = 0.5$, $\{e_t\}$ is a weakly stationary martingale difference sequence whereas this is not the case for $\lambda = 1$. Nevertheless the μ_t sequence may still, at least in the case where $q = 0$ and $p = 1$, be a stationary first order Markov process.

2. It defines the mean as a distributed lag in the previous regression terms. The expected value of the mean function will not be the current deterministic part of the process mean, even approximately, which might make interpretation of the coefficients β difficult for practical use.

3. The model implies that there must always be at least one non-trivial lag term in the autoregressive component. Our model does not require this and allows pure MA, pure AR and mixed ARMA models.

3.3 Properties of the Process

Note that $\{e_t : t \leq s-1\}$, given initial conditions $\{\mu_t : -p+1 \leq t \leq 0\}$ is equivalent to $\{Y_t : t \leq s-1\}$ or $\{\mu_t : t \leq s-1\}$ from which is follows that the e_t form a martingale difference sequence since

$$E(e_s|\mathcal{F}_{s-1}^e) = 0$$

where $\mathcal{F}_{s-1}^e$ is the σ-algebra generated by $\{e_t : t \leq s-1\}$. Hence the e_t have zero mean and variance

$$E(e_t^2) = E[E(e_t^2|\mu_t)] = \mu_t^{1-2\lambda}$$

which, for $\lambda = 0.5$, is unity. Also, from the martingale difference property, the covariance, for $s \neq t$, is

$$E(e_t e_s) = 0.$$

From the above properties we have, for any λ,

$$E(W_t) = \mathbf{x}_t^T\boldsymbol{\beta}$$

which is a desirable property for the log mean. Compare with the first desiderata listed above. Also

$$\mathrm{Var}(W_t) = \sum_{i=1}^{p}\gamma_i^2\mu_{t-i}^{1-2\lambda}$$

and for $s = t + l$ for $l > 0$

$$\mathrm{Cov}(W_t, W_s) = \sum_{i=1}^{p-l}\gamma_i\gamma_{i+l}\mu_{t-i}^{1-2\lambda}$$

and again, if $\lambda = 0.5$ the covariances do not depend on time t.

Because of the above the case where $\lambda = 0.5$ seems to have nice properties. We consider this case in depth now.

Now we know that the W_t have mean exactly equal to the linear predictor. However it does not follow that the μ_t have mean equal to the e. However we have

$$\begin{aligned} W_t &= \mathbf{x}_t^T\boldsymbol{\beta} + U_t \\ &\approx \mathbf{x}_t^T\boldsymbol{\beta} + U_t' \end{aligned}$$

in the sense that the distributions will be similar where U_t' is a Gaussian stationary sequence with mean zero and variances and covariances matched to those for U_t. Roughly speaking U_t' is a latent process. Hence, using the results obtained for latent processes we have, again in the case $\lambda = 0.5$,

$$\begin{aligned} E(e^{W_t}) &\approx e^{\mathbf{x}_t^T\boldsymbol{\beta} + \frac{1}{2}\mathrm{Var}(U_t)} \\ &= e^{\mathbf{x}_t^T\boldsymbol{\beta} + \frac{1}{2}\sum_{i=1}^{p}\gamma_i^2}, \end{aligned}$$

so that, in practice the bias of $E(\mu_t)$ as an estimate of $e^{\mathbf{x}_t^T\boldsymbol{\beta}}$ can be approximately adjusted for and, perhaps most importantly the interpretation of the regression coefficients, other than the intercept or constant term, are (at least to the level of approximation) interpretable as the amount by which the mean of Y_t would change for a unit change in the regressors.

While the distribution of e_t is not normally distributed the linear combination $U_t = \sum_{i=1}^{p}\gamma_i e_{t-i}$ will have a distribution more closely well approximated by a sequence of correlated normal random variables. The extent to which the joint distribution of the sequence $\{e_t\}$ differs from a process of independent Gaussian random variables with zero mean and unit variance will govern the extent to which the approximation

$$E(e^{W_t}) = e^{\mathbf{x}_t^T\boldsymbol{\beta} + \frac{1}{2}\sum_{i=1}^{p}\gamma_i^2}$$

holds.

Overall, the $\lambda = 0.5$ has a number of desirable properties:

1. It is easily interpretable on the linear predictor scale *and* on the scale of the mean μ_t with the regression parameters (approximately) directly interpretable as the amount by which the mean of the count process at time t will change for a unit change in the regressor variable.

2. An approximately unbiased plot of the μ_t can be generated by

$$\hat{\mu}_t = \exp(\hat{W}_t - \frac{1}{2}\sum_{i=1}^{p}\hat{\gamma}_i^2)$$

   where estimates have been used throughout.

3. The model is easy to use for prediction. In fact $\hat{\mu}_t$ should be used as the one step ahead forecast of Y_t, given a value for $\mathbf{x}_t$ or a reliable forecast of it.

4. The model provides a mechanism for adjusting the inference about the regression parameter $\boldsymbol{\beta}$ for a form of serial dependence.

5. The model is generalizable to "autoregressive" type lag structure and mixed autoregressive-moving average structure.

3.4 Estimation and Inference

3.4.1 Estimation

In this section we will only consider estimation and inference properties for the observation-driven model of Section 3.2 when $\lambda = 0.5$. It might be thought to be reasonable to use an iterated GLM procedure in which at each stage the values of the sequence $\{e_t\}$ are used in a standard GLM procedure to obtain estimates of β and $\xi = (\phi^T, \gamma^T)^T$ which are then used to redefine $\{e_t\}$ and so on iteratively. This does not appear to converge to the values which maximize the likelihood. Furthermore, and perhaps more importantly, the standard errors obtained will not be correct for β and ξ using this iterated method. However computation of the likelihood and its first and second derivative is not difficult and can be done recursively using a simple Newton-Raphson update procedure. Correct standard errors are then also available.

Ignoring terms which do not involve the parameters, the log-likelihood, as a function of $\theta = (\beta^T, \xi^T)^T$, is given by

$$L(\theta) = \sum_{t=1}^{n} [Y_t W_t(\theta) - e^{W_t(\theta)}],$$

where

$$\log(\mu_t) = W_t(\theta) = \mathbf{x}_t^T \beta + \sum_{i=1}^{\infty} \tau_i(\xi) e_{t-i}(\theta) \tag{3.4}$$

and

$$e_t = (Y_t - \mu_t)/\sqrt{\mu_t}.$$

First and second derivatives are given by the following expressions

$$\frac{\partial L}{\partial \theta} = \sum_{t=1}^{n} (Y_t - \mu_t) \frac{\partial W_t}{\partial \theta} = \sum_{t=1}^{n} e_t \mu_t^{1/2} \frac{\partial W_t}{\partial \theta}$$

and

$$\begin{aligned} \frac{\partial^2 L}{\partial \theta \partial \theta^T} &= \sum_{t=1}^{n} [(Y_t - \mu_t) \frac{\partial^2 W_t}{\partial \theta \partial \theta^T} - \mu_t \frac{\partial W_t}{\partial \theta} \frac{\partial W_t}{\partial \theta^T}] \\ &= \sum_{t=1}^{n} [e_t \mu_t^{1/2} \frac{\partial^2 W_t}{\partial \theta \partial \theta^T} - \mu_t \frac{\partial W_t}{\partial \theta} \frac{\partial W_t}{\partial \theta^T}]. \end{aligned}$$

In order to calculate these, recursive expressions for $\frac{\partial W_t}{\partial \theta}$ and $\frac{\partial^2 W_t}{\partial \theta \partial \theta^T}$ can readily be derived and programmed. Thus calculation of the likelihood and its first and second derivatives (from which the standard errors are derived - see below) can be implemented in a straightforward fashion. We have found that implementation in Splus on a PC provides reasonably fast computation of estimates based on the Newton-Raphson method.

3.4.2 Initial Estimates

To initialize the Newton-Raphson recursions we have found that using the GLM estimates without terms for the autoregressive moving average terms together with zero initial values for the ARMA terms gives reasonable starting values. Convergence in all cases reported below (in which the first derivatives were less than 10^{-8}) occurred within at most 6 iterations from these starting conditions.

3.4.3 Asymptotics

It would appear that the asymptotic properties of the maximum likelihood estimates considered in the previous section are straightforward to establish due primarily to the martingale difference properties of $\{e_t\}$. Nevertheless, a formal proof of these asymptotic properties is difficult without a more detailed understanding on the underlying distributional structure of the $\{e_t\}$ sequence. Assuming that the regressor sequences $x_{j,t}$ satisfy the conditions stated in Section 2.2 above and that

$$\Omega^{-1} = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} e^{W_t(\boldsymbol{\theta}_0)} W_t'(\boldsymbol{\theta}_0) W_t'(\boldsymbol{\theta}_0)^T$$

exists and is non-singular, we conjecture that

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \Omega^{-1}).$$

Asymptotic standard errors of the estimates are then easily obtained from

$$\hat{\Omega} = -\left(\frac{1}{n} \frac{\partial^2 L(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right)^{-1}$$

as

$$\hat{\sigma}_{\hat{\boldsymbol{\theta}},j} = \sqrt{\frac{\hat{\Omega}_{jj}}{n}}.$$

4 Examples

4.1 Analysis of the Polio Data

We provide an analysis of the polio data. These data have been much analyzed (list of references) but all these analyses appear to provided different inferences especially about the important issue of trend. A summary is provided below for reference.

Study	Trend ($\hat{\beta}$)	s.e.($\hat{\beta}$)	t-statistic
GLM Estimate	-4.80	1.40	-3.43
Zeger (1988)	-4.35	2.68	-1.62
Chan & Ledolter (1995)	-4.62	1.38	-3.35
Kuk & Cheng (1996) MCNR	-3.79	2.95	-1.28
Jorgensen et al. (1995)	-1.61	0.018	-89.7
Fahrmeir and Tutz (1994) 5 lags in Y	-3.33	2.00	-1.67

Clearly there is lack of agreement on the values of the parameter estimates and the standard errors that should be attached to them. Some of these standard errors appear to be unreasonably small in the light of the other results and those obtained below using the observation driven model.

We fit model (3.4) to the data with the same regression terms as used in Zeger (1988) (and centered in an identical way as described there). Two models were considered. The first started with $p = 6$ (the first 6 moving average terms) and $q = 0$ (no autoregressive components). The second started with $p = 0$ and $q = 6$. In both cases any autoregressive or moving average terms which were not significant at the 5% level were dropped from the model and the model refitted to arrive at the following table of results.

Model Term	AR $\hat{\theta}$	AR $\hat{\sigma}^{(1)}_{\hat{\theta}}$	MA $\hat{\theta}$	MA $\hat{\sigma}^{(2)}_{\hat{\theta}}$
$\hat{\beta}_0$ (Intercept)	.138	.117	.130	.114
$\hat{\beta}_1$ (Trend)	-3.83	2.26	-3.93	2.18
$\hat{\beta}_2$ (annual cosine)	-.099	.105	-.099	.118
$\hat{\beta}_3$ (annual sine)	-.506	.128	-.531	.141
$\hat{\beta}_4$ (semi-annual cosine)	.230	.127	.211	.117
$\hat{\beta}_5$ (semi-annual sine)	-.397	.123	-.393	.116
$\hat{\phi}_1$ or $\hat{\gamma}_1$	.227	.053	.218	.056
$\hat{\gamma}_2$			.127	.046
$\hat{\phi}_5$ or $\hat{\gamma}_5$	.105	.050	.087	.043
$L(\hat{\beta}, \hat{\xi})$ (log-likelihood)	-119.6		-118.9	
$\hat{\beta}_0 + \frac{1}{2}\sum \hat{\gamma}_i^2$ (Adj'd I'cept)	-		0.166	

The autocorrelations of the e_t using Zeger's method and using the optimal estimation method do not appear to be significant at any lag up to lag 12 (the largest values estimated). Their mean is 0.0186 and variance is 1.5015. When the largest 5 values of the e_t, all of which exceed 3 are excluded, the mean is -0.096 with variance 1.099. It would seem as though the values of the process at times $t = 7, 34, 35, 74, 113$ which have Pearson (predictive) residuals e_t which indicates they could be outliers.

Use of the Q statistic to determine if a latent process is present gives a value of 1.64 (1.99 adjusted for bias due to fitting parameters) which is significant at the 5% level on a 1 sided test (at the 2.3% level using the bias adjusted version). Note that the distributional properties of the alternative statistics S_a and $\tilde{Q}$ are not established for these lagged regression models. It is likely that the simpler statistic Q will be conservative, however, so the marginally significant result obtained here indicates that not all the additional variation in the observed counts is accounted for be the lagged regression model

4.2 Analysis of the UK SIDS Data

Campbell (1994) considers the time series of sudden infant death syndrome cases in the UK with the objective of understanding the relationship between temperature fluctuations and the number of daily SIDS deaths. On the basis of a model which does not include temperature he reasons that the method of Zeger is required to adjust for serial dependence in the observed counts. He then applies Zeger's (1988) method to a model in which temperature effects are included.

In our view there is no need to adjust for serial dependence in the full model (which includes temperature effects). In fact when we analyze these data we find that evidence for serial dependence through a latent process is non-existent. Using the model numbers that Campbell uses we find, upon application of the tests for the existence of a latent process the following test statistic values.

Model	S_a test	ACF test H^2_{OPT}, $L = 20$
Model 1	3.07	241.2
Model 2	1.60	109.1
Model 3	1.44	79.0
Model 4	0.58	12.30

Note that when a linear trend effect is included (model 4) there is no evidence of the existence of a latent process.

We also examined the evidence for serial dependence using an observation driven model. Several ARMA terms were added to model 4 including an ARMA(1,0), ARMA((1,7),0), ARMA(0,1). In none of these models was there any evidence of significant effects, all estimated ARMA coefficients being not significant at any where near the 5% level.

In conclusion, application of methods for detecting a latent process or for fitting an observation driven model leads to the conclusion that, provided the full set of regression terms (including temperature most importantly), there is no evidence of serial dependence additional to that in the regressor terms.

4.3 Analysis of the Asthma Data

Exploratory analysis of these data suggested the need to include model terms for a Sunday effect, a Monday effect, a possible increasing linear trend in time and Fourier series terms to model the seasonal pattern in the data. It was found necessary to include terms for $\cos(2\pi kt/365)$ and $\sin(2\pi kt/365)$ for $k = 1, 2, 3, 4$. Initial estimates of regression parameters were obtained using a Poisson regression. The residuals from this fit indicated that the there was significant serial dependence in the observed asthma counts. Indeed application of the test statistics described in Davis, Dunsmuir and Wang (1997) gave the following results. The Q statistic was observed to be 2.41 and the variance of the residuals was 0.72 with standard error of 0.024 indicating significant additional variance than that explained by the Poisson regression alone. The portmanteau test for serial dependence was observed to be 46.7 on 20 df which is highly significant (P = .00064) also indicating serial dependence. Significant values in the autocorrelation estimates (based on Zeger's method) were observed at lags 1,2,3,5,7,10. The trend term was not significant in the Poisson regression and also in the observation driven model to be described next. Accordingly details of its estimation will not be given and it was omitted from the model. Thus the asthma presentations observed at Cambelltown appear not to have a significant increasing trend over the 4 years of observations. All Fourier terms had at least one significant coefficient of each pair. These were retained in the model to be described next.

To model the observed serial dependence we used the GLMARMA model with lags 1,2,3,5,7,10 for the AR component and no moving average component, at least initially. The resulting estimates are given in the table displayed below.

The log-likelihood value is -776.22 for this model indicating substantial improvement in the fit. Note that the AR coefficients at lags 2 and 5 are not significant. The model was refit dropping the AR coefficients at these two lags. The results are also shown in the above table for ease of comparison with the previous model. Note that the log-likelihood is now -778.2398 which shows a non-significant change from the model with the additional AR terms at lags 2 and 5.

Application of the various test statistics to the residuals (although probably not valid for the lagged model) are as follows. The Q statistic was now observed to be 1.75 and the variance of the residuals was 0.052 with standard error of 0.024 still indicating significant additional variance than that explained by the Poisson regression alone. The portmanteau test for serial dependence was observed to be 18.0 on 20 df which is not significant. The lag 2 autocorrelation is significantly different from zero at about the 1% level. No other significant values in the autocorrelation estimates (based on Zeger's method) were observed.

In summary there remains some slight overdispersion not explained by the lagged AR component of this observation driven model.

An alternative is to model the lag dependence using a moving average model. The most parsimonious MA is MA(1,3) with log-likelihood value -786.87, but this is not an improvement over the AR. On the other hand the MA model is easier to use in practice involving a finite linear

combination of past "errors" so that if it is not a noticeable reduction in fit it might be preferred.

	AR(1,2,3,5,7,10)		AR(1,3,7,10)	
Term	Parameter	se	Parameter	se
(Intercept)	0.533	0.031	.532	.030
Sunday effect	0.233	0.054	.240	.054
Monday effect	0.245	0.054	.244	.054
$\cos(2\pi t/365)$	-0.163	0.039	-.163	.037
$\sin(2\pi t/365)$	0.360	0.039	.362	.036
$\cos(4\pi t/365)$	-0.066	0.039	-.067	.038
$\sin(4\pi t/365)$	0.021	0.039	.021	.035
$\cos(6\pi t/365)$	-0.080	0.038	-.080	.036
$\sin(6\pi t/365)$	0.008	0.038	.009	.036
$\cos(8\pi t/365)$	-0.148	0.038	-.152	.036
$\sin(8\pi t/365)$	-0.057	0.038	-.057	.035
ϕ_1	0.044	0.017	.047	.017
ϕ_2	0.026	0.018	-	
ϕ_3	0.046	0.018	.049	.017
ϕ_5	0.023	0.018	-	-
ϕ_7	0.058	0.017	.059	.017
ϕ_{10}	0.038	0.018	.041	.018

References

[1] Brannas, K. and Johansson, P. (1994) "Time series count data regression", Commun. Statist.-Theory and Meth., 23 (10), 2907-2925.

[2] Brockwell, P. J. and Davis, R. A. (1991) Time Series: Theory and Methods (2nd edition), Springer-Verlag, New York.

[3] Brockwell, P. J. and Davis, R. A. (1996) Introduction to Time Series and Forecasting, Springer, New York.

[4] Campbell, M. J. (1994) "Time series regression for counts: an investigation into the relationship between sudden infant death syndrome and environmental temperature", J. R. Statist. Soc. A, 157, 191-208.

[5] Carlin, B. P., Polson, N. G. and Stoffer, D. S. (1992) "A Monte-Carlo approach to nonnormal and nonlinear state-space models", JASA, 87, 493-500.

[6] Chan, K. S. and Ledolter, J. (1995) "Monte Carlo EM estimation for time series models involving counts", JASA, 90, 242-252.

[7] Davis, M. H. A. and Vinter, R. (1985) Stochastic Modeling and Control, Chapman & Hall, New York.

[8] Davis, R. A. and Dunsmuir, W. T. M. (1997) "A new observation driven model for Poisson time series". In preparation.

[9] Davis, R. A., Dunsmuir, W. T. M. and Wang, Y. (1997) "Detecting autocorrelation in the mean of a Poisson regression". In preparation.

[10] Dean, C. B. (1992) "Testing for overdispersion in Poisson and binomial regression models", JASA, 87, 451-457.

[11] Dean, C. and Lawless J. F. (1989) "Tests for detecting overdispersion in Poisson regression models", JASA, 84, 467-472.

[12] Diggle, P. J., Liang, K-Y and Zeger, S. L. (1994) Analysis of Longitudinal Data, Oxford University Press, Oxford.

[13] Durbin, J. and Koopman, S. J. (1997) "Monte Carlo maximum likelihood estimation for non-Gaussian state space models", Biometrika (to appear).

[14] Fahrmeir, L. and Kaufmann, H. (1987) "Regression models for non-stationary categorical time series", J. Time Ser. Anal., 8, 147-160.

[15] Fahrmeir, L. and Tutz, G. (1994) Multivariate Statistical Modeling Based on Generalized Linear Models, Springer-Verlag, New York.

[16] Hannan, E. J. (1970) Multiple Time Series, John Wiley & Sons, New York.

[17] Harvey, A. C. and Fernandes, C. (1989) "Time series models for count or qualitative observations", J. Business and Economic Statistics, 7, 407-17.

[18] Hannan, E. J. and Deistler, M. (1988) The Statistical Theory of Linear Systems, John Wiley & Sons, New York.

[19] Jorgensen, B., Lundbye-Christensen, S., Song, X.-K. and Sun, L. (1995) "A state space model for multivariate longitudinal count data", Technique Report 148, Department of Statistics, University of British Columbia.

[20] Judge, G. G. (1985) The Theory and Practice of Econometrics (2nd edition), John Wiley & Sons, New York.

[21] Kuk, A. Y. C. and Cheng, Y. W. (1996) "The Monte Carlo Newton Raphson algorithm", Working Paper, Department of Statistics, UNSW.

[22] Saldiva, P. H., Pope, C. A. 3rd, Schwartz, J. and Dockery, D. W. (1995) "Air pollution and mortality in elderly people: a time-series study in Sao Paulo, Brazil", Archives of Environmental Health, 50, 159-163.

[23] Schwartz, J., Spix, C. and Touloumi G. (1996) "Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions", Journal of Epidemiology and Community Health, 50 (Supplement), S3-S11.

[24] Shephard, N. (1995) "Generalized linear autoregressions", Working Paper, Nuffield College, Oxford.

[25] Zeger, S. L. (1988) "A regression model for time series of counts", Biometrika, 75, 621-629.

[26] Zeger, S. L. and Qaqish, B. (1988) "Markov regression models for time series: a quasi-likelihood approach", Biometrics, 44, 1019-1031.