Best Practice Recommendation for Forecasting Counts

Brajendra C. Sutradhar
Department of Mathematics and Statistics, Memorial University of Newfoundland
St. John's, NL, Canada A1C 5S7

Abstract

The existing techniques of forecasting a future count either treat the time series of counts as a Gaussian time series or use a random effects based dynamic Poisson model. The normality based approach may not yield valid forecasting, whereas the random effects based model usually generates a complex correlation structure for the time series of counts which may be impractical to use for forecasting. Moreover, when the time series contains moderately large or large counts, the later random effects based models are known to be inefficient in forecasting a future count, based on such a time series of large counts.

In this report, we propose an observation driven non-stationary correlation model to fit a time series of counts with possible overdispersion. Analogoues to the Gaussian time series techniques, we develop forecasting functions to forecast future counts. The proposed forecasting approach is simpler than the existing approaches and it is shown through a simulation study that this approach provides satisfactory forecasting for a future count, irrespective of the cases whether time series contains small or large counts. Thus, the proposed approach may be recommended as the best practical approach for forecasting a count. The forecasting methodology is illustrated by analyzing a U.S. time series of polio counts.

Some key words: Consistency; Efficiency; Latent process driven longitudinal correlation; Observations driven longitudinal autocorrelation; Overdispersion; Regression effects; Forecasting.

1 Introduction

Forecasting counts is an important reserach topic in many socio-economic sectors. For example, forecasting the number of polio patients for a state/country

based on the time series of polio counts is an important problem for health economics. Similarly, forecasting the number of tourists for a city/country and forecasting the number of patents to be awarded to a firm are important economic problems. The modelling of the time series of counts, in particular the non-stationary time series of counts, is however not easy. This is mainly because of the difficulty of writing the multivariate distribution for the correlated counts recorded over the years. This hampers the forecasting ability naturally.

There exist some studies where the time series of counts are treated as Gaussian time series and normality based available forecasting techniques are used (see for example, Kulendran and King (1997)) to forecast a future count. As in this approach forecasting is made without challenging the Gaussian assumption for the Poisson or negative binomial data, the forecasting may not be valid. Further, there exist a random effects based dynamic modelling approach (Zeger (1988), Harvey and Farnandes (1989), Settimi and Smith (2000)). In this approach, random parameters defined at a time point t are assumed to have a functional relationship with parameters defined for past times, with the initial random parameter having a suitable probability distribution. To be specific, Zeger (1988) has modelled the time series of counts by assuming that each of the count responses is affected by a specific random effect. If the individual random effect follows a suitable gamma distribution, then the corresponding count response will have a negative binomial distribution. As far as the joint distribution is concerned, it is reasonable to assume that conditional on the random effects, the count responses follow independent Poisson distributions. Next by assuming that the random effects are correlated with a Gaussian type auto-correlation structure, Zeger (1988) has developed a unconditional correlation structure which is not easy to estimate as this structure is dependent on the unknown correlation structure of the random effects. Furthermore, forecasting aspects of the data were not considered. Similarly to Zeger (1988), Harvey and Fernandes (1989) have also modelled the time series of counts by using Poisson distributions for the count responses conditional on the random effects. They used a suitable dynamic relationship among the random effects which in turn make the count responses correlated. Note that their approach generates a negative binomial distribution for the random count response y_t at time t, say, conditional on the available past count $Y_{t-1} = y_{t-1}$, whereas Zeger's model produces marginal negative binomial distribution for each y_t. Consequently, even

though one can develop a forecasting function based on the conditional distribution approach of Harvey and Fernandes (1989), it seems more appropriate to develop suitable conditional distribution for forecasting where the responses will have marginal negative binomial distribution. Recently, Settimi and Smith (2000) used a Bayesian approach, which is similar to Harvey and Fernandes's (1989) approach, for forecasting for discrete time series data. More recently, Davis et al (2003) considered a dynamic conditional Poisson probability model which may be used for forecasting the future counts but their correlation model for the random effects appear to be arbitrary.

As opposed to the random effects based models for the time series of counts, there also exist observation driven models. For example, McKenzie (1986, 1988) [see aslo Al-Osh and Alzaid (1987)] introduced auto-regressive models for stationary Poisson and negative binomial data. Recently, following McKenzie, Jowaheer and Sutradhar (2002) used the observations driven stationary negative binomial correlation model to analyze non-stationary negative binomial data in the longitudinal set up. More recently, Freeland and McCabe (2004) used the stationary correlation models for Poisson data in the context of forecasting future counts. Note however that as in practice one frequently encounters non-stationary count data, it raises the concern to use the appropriate correlation models for non-stationary data for the purpose of forecasting. Moreover, the non-stationary count data may also exhibit overdispersion so that a non-stationary negative binomial model may be more appropriate than non-stationary Poisson model.

In this report, we consider observation driven Poisson as well as negative binomial correlation model for non-stationary count data. The forecasting functions based on Poisson and negative binomial correlation models are constructed and their forecasting performances are examined through a simulation study. The proposed forecasting methodology is illustrated by forecasting future counts for the U.S. polio count data.

2 Forecasting Based on Existing Models For Time Series of Counts

2.1 Correlated Random Effects Based Marginal Models

Let y_t $(t = 1, \cdots, T)$ be the count response recorded at time t and x_t $(t = 1, \cdots, T)$ be the corresponding $p \times 1$ vector of covariates. Further let $\beta = (\beta_1, \cdots, \beta_p)^{'}$ be the p-dimensional vector of regression effects. Zeger (1988) and Davis et al (2000) have considered a sequence of correlated random effects $\{\theta_t\}$ such that conditional on θ_t, $y_t \sim P(\mu_t^*)$, i.e., y_t has the Poisson distribution given by

$$f(y_t \mid \theta_t) = \frac{e^{-\mu_t^*} \mu_t^{* y_t}}{y_t!}, \tag{2.1}$$

where $\mu_t^* = \mu_t \theta_t$ with $\mu_t = e^{x_t^{'} \beta}$ and $\theta_t = e^{\gamma_t}$ (say), so that $E(Y_t \mid \theta_t) = var(Y_t \mid \theta_t) = \mu_t^*$. Next by using

$$E(\theta_t) = 1, var(\theta_t) = c > 0, \rho_\theta(l) = \alpha^{-1} E(\theta_t - 1) E(\theta_{t+l} - 1) \tag{2.2}$$

they modelled the unconditional means, variances and correlations of the observations with

$$E(Y_t) = \mu_t, var(Y_t) = \mu_t + c\mu_t^2$$

$$\rho_y(l) = \frac{\rho_\theta(l)}{\{1 + (c\mu_t)^{-1}\}^{\frac{1}{2}} \{1 + (c\mu_{t+l})^{-1}\}^{\frac{1}{2}}} \tag{2.3}$$

It is clear from (2.3) that $\rho_y(l)$ heavily depends on the correlations $\rho_\theta(l)$ of the random effects. Consequently, this approach of modelling the correlations of the observations has some major limitations. For example, in practice, it is almost impossible to identify the correlations of the random effects which hampers the estimation of the correlations of the count observations. To be specific, even if one estimates $\rho_y(\ell)$ by the method of moments using directly the sample lag ℓ auto-correlation, in some cases it becomes impossible to know whether this is a valid estimate. This is because, under this approach, the range of lag correlation of the responses depends on the value of the corresponding lag correlation of the random effects which is however unknown. To make it clear, suppose that the random effect θ_t marginally follow the log-normal distribution, i.e., $\gamma_t \sim N(-\sigma^2/2, \sigma^2)$ and $p = 1$. It then follows that

$$\rho_y(l) = \frac{e^{\tau_\gamma(l)} - 1}{\mu^{-1} + (e^{\sigma^2 - 1})},$$

where $\tau_\gamma(l) = cov(\gamma_t, \gamma_{t+l})$ and $\mu = e^{\beta_1}$. Consequently, it can be shown [Davis et al (2000)] that

$$0 \leq \rho_y(l) \leq \frac{e^{\tau_\gamma(l)} - 1}{e^{\sigma^2 - 1}} \leq \frac{\tau_\gamma(l)}{\sigma^2} = \rho_\gamma(l). \tag{2.4}$$

The above range relationship (2.4) indicates that when one estimate $\rho_y(l)$ by sample lag l correlation $\hat{\rho}_y(l)$, it is not possible to know whether it is a valid estimate until one knows $\rho_\gamma(l)$. But $\rho_\gamma(l)$ is not known in practice. Furthermore, it is not easy to interpret such correlations defined in (2.3) of the observations, whereas in the Gaussian time series set up, $\rho_y(l)$ has closed form expressions and they are easy to interpret.

2.1.1 Forecasting

Zeger (1988) and Davis et all (2000) did not discuss the forecasting aspects. In fact, as this is a marginal approach, finding the conditional mean of y_t for given y_{t-1} does not appear to be easy. To be specific, to find the conditional mean, i.e., the forecasting function, one will require the joint distribution of the correlated random effects $\theta_1, \ldots, \theta_t, \ldots, \theta_T$. Let $g(\theta_1, \ldots, \theta_T)$ be the joint density of these random effects so that θ_t marginally has the gamma distribution with mean 1 and variance $c = 1/c_1$, that is,

$$g_t(\theta_t) = \{{c_1}^{c_1}/\Gamma(c_1)\}{\theta_t}^{c_1 - 1} e^{(-c_1\theta_t)}.$$

Note that finding such a joint density function is, however, not easy. If $g(.)$ is available, one then finds the joint distribution of y_t and y_{t-1} as

$$f(y_t, y_{t-1}) = \int_{\theta_1} \ldots \int_{\theta_t} \ldots \int_{\theta_T} \frac{e^{-(\mu_t^* + \mu_{t-1}^*)} {\mu_t^*}^{y_t} {\mu_{t-1}^*}^{y_{t-1}}}{y_t! y_{t-1}!} g(\theta_1, \ldots, \theta_t, \ldots, \theta_T) \partial\theta_1 \ldots \partial\theta_t \ldots \partial\theta_T, \tag{2.5}$$

and compute the forecasting function $E(Y_t|y_{t-1})$ from the conditional distribution

$$f(y_t|y_{t-1}) = f(y_t, y_{t-1})/f_1(y_{t-1})$$

where the marginal density of y_{t-1} is obtained from (2.1) as

$$f_1(y_{t-1}) = \int_{\theta_{t-1}} f(y_{t-1}|\theta_{t-1}) g_{t-1}(\theta_{t-1}) \partial\theta_{t-1}.$$

It is clear from the above description that in this random effects based marginal approach, the computation for the forecasting function is extremely difficult.

Consequently, we do not pursue this procedure any further for the purpose of forecasting.

2.2 Dynamic Models Through Past Observations Based Random Effects

In this approach, similar to Zeger (1988), y_t conditional on θ_t still has the distribution given by (2.1), i.e.,

$$f(y_t \mid \theta_t) = \frac{e^{-\mu_t^*}\mu_t^{*y_t}}{y_t!},$$

with $\mu_t^* = \mu_t\theta_t$ with $\mu_t = e^{x_t'\beta}$ and $\theta_t = e^{\gamma_t}$. As far as the distribution of θ_t is concerned, Harvey and Fernandes (1989) [see also Settimi and Smith (2000, p. 139-140)] assumed that the random effect θ_t conditional on the past count response y_{t-1} follows a gamma distribution $G(a_t,\ b_t)$ given by

$$g\left(\theta_t \mid y_{t-1}\right) = \frac{e^{-b_t\theta_t}\theta_t^{a_t-1}}{\Gamma(a_t)b_t^{-a_t}}, \theta_t > 0, \tag{2.6}$$

whereas the random effects have the marginal distributions given by

$$g\left(\theta_{t-1} \mid y_{t-1}\right) = \frac{e^{-b_t^*\theta_{t-1}}\theta_{t-1}^{a_t^*-1}}{\Gamma(a_t^*)b_t^{*-a_t^*}}, \theta_{t-1} > 0, \tag{2.7}$$

such that $a_t = wa_{t-1}^*$ and $b_t = wb_{t-1}^*$, where $w > 0$ is a scale parameter.

It then follows that

$$\begin{aligned}
f\left(y_t \mid y_{t-1}\right) &= \int_0^\infty f\left(y_t \mid \theta_t\right) g\left(\theta_t \mid y_{t-1}\right) d\theta_t \\
&= \int_0^\infty \frac{e^{-\mu_t\theta_t}(\mu_t\theta_t)^{y_t}}{y_t!}\frac{e^{-b_t\theta_t}\theta_t^{a_t-1}}{\Gamma(a_t)b_t^{-a_t}}d\theta_t \\
&= \frac{b_t^{a_t}{\mu_t}^{y_t}}{\Gamma(a_t)y_t!}\frac{\Gamma(y_t+a_t)}{(b_t+\mu_t)^{y_t+a_t}} \\
&= \frac{\Gamma(y_t+a_t)}{\Gamma(a_t)y_t!}\left(\frac{b_t}{\mu_t}\right)^{a_t}\left(1+\frac{b_t}{\mu_t}\right)^{-(y_t+a_t)} \\
&= \frac{\Gamma(y_t+a_t)}{\Gamma(a_t)y_t!}\left(\frac{1}{1+\frac{b_t}{\mu_t}}\right)^{y_t}\left(1-\frac{1}{1+\frac{b_t}{\mu_t}}\right)^{a_t}. \qquad (2.8)
\end{aligned}$$

The form stated in (2.8) is matched with that in Jowaheer and Sutradhar (2002) and accordingly denoted by $NB\left(a_t, \frac{\mu_t}{b_t}\right)$ which may be used for forecasting lag 1 future count. More specifically, lag 1 forecasting function is given by

$$E\left(Y_t \mid y_{t-1}\right) = \frac{a_t}{b_t}\mu_t = \frac{a^*_{t-1}}{b^*_{t-1}}\mu_t = r^*_t \tag{2.9}$$

with its variance as

$$V\left(Y_t \mid y_{t-1}\right) = r^*_t + \frac{1}{a_t}{r^*_t}^2 = \frac{a^*_{t-1}}{b^*_{t-1}}\mu_t + \frac{1}{w}\frac{a^*_{t-1}}{{b^*_{t-1}}^2}\mu^2_t, \tag{2.10}$$

implying that w is an overdispersion parameter.

2.2.1 Estimation of Parameters

Note that under the assumption that the initial response y_0 follows a suitable distribution, say $f^*(y_0)$, one may derive the exact likelihood as

$$L = f^*_(y_0)f(y_1|y_0)f(y_2|y_1, y_0)\ldots f(y_T|y_{T-1}, \ldots, y_1, y_0), \tag{2.11}$$

which requires the modelling of $f(y_t|y_{t-1}, \ldots, y_1, y_0)$. It is clear from (2.5) that Harvey and Fernandes (1989) avoided the modelling for this general conditional distribution, by considering a lag 1 type dependence of θ_t on y_{t-1}. This yields the log-likelihood function given by

$$\begin{aligned}
\mathcal{L}(w, \beta) &= log\prod_{t=1}^{T} f\left(y_t \mid y_{t-1}\right) \\
&= \sum\Bigg[log\Gamma(y_t + a_t) - log\Gamma(a_t) - log(y_t!) + a_t log\left(\frac{b_t}{\mu_t}\right) \\
&\quad - (y_t + a_t)log\left(1 + \frac{b_t}{\mu_t}\right)\Bigg]
\end{aligned} \tag{2.12}$$

which must be maximized to estimate the parameters involved. To be specific, using the recurrence relation $a_t = wa^*_{t-1}$ and $b_t = wb^*_{t-1}$ such that $a^*_0 = b^*_0 = 0$, the log-likelihood function (2.12) was maximized by Harvey and Fernandes (1989) to estimate w and β.

2.2.2 Predictive Distribution and Forecasting:

Note that

$$\theta_t \mid y_{t-1} \sim G\left(a_t, b_t\right),$$

where $a_t = wa^*_{t-1}$ and $b_t = wb^*_{t-1}$. When y_t becomes available, the predictive distribution of θ_t given y_t is obtained by using the well-known Bayesian approach. More specifically it can be shown that

$$\begin{aligned} g(\theta_t|y_t) &= [f(y_t|\theta_t)g(\theta_t|y_{t-1})]/f(y_t|y_{t-1}) \\ &= \frac{e^{-\theta_t(\mu_t+b_t)}{\theta_t}^{y_t+a_t-1}}{\Gamma(y_t+a_t)(\mu_t+b_t)^{-(y_t+a_t)}}, \end{aligned} \tag{2.13}$$

implying that

$$\theta_t \mid y_t \sim G\left(a^*_t, b^*_t\right),$$

where $a^*_t = a_t + y_t$ and $b^*_t = \mu_t + b_t$, with $a_t = wa^*_{t-1}$, and $b_t = wb^*_{t-1}$.

Note that it follows by (2.6) and (2.7) that

$$\begin{aligned} E\left(Y_{t+1} \mid y_t\right) &= \mu_{t+1}E(\theta_{t+1}|y_t) \\ &= \mu_{t+1}E(\theta_t|y_t) \\ &= \mu_{t+1}\frac{a^*_t}{b^*_t} = \mu_{t+1}[y_t + wa^*_{t-1}]/[\mu_t + wb^*_t] \\ &= \mu_{t+1}[\sum_{j=0}^{t-1} w^j y_{t-j}]/[\sum_{j=0}^{t-1} w^j \mu_{t-j}] = r^*_{t+1} \end{aligned} \tag{2.14}$$

with its variance as

$$V\left(Y_{t+1} \mid y_t\right) = r^*_{t+1} + \frac{1}{a^*_{t+1}}{r^*_{t+1}}^2. \tag{2.15}$$

2.2.3 Forecasting Performance : An Application to the U.S Polio Count Data

Forecasting counts in biomedical science is an important problem for future health planning. Zeger (1988) and Davis et al (2000) analyzed a time series of counts (see Figure 1) on the the monthly number of cases of poliomyelitis reported by the U.S. Centers for Disease Control for the years 1970-1983. Here total number of observations is $T = 168$. These authors however did not consider the forecasting issues, rather, they dealt with modelling the data and estimation of the parameters of the model. Now to examine the performance of the forecasting function (2.14) due to Harvey and Fernandes (1989), we have decided to fit their model to this data set of length say $T = 160$ and then

forecast the count for time points $T+1=161$, in order to see whether the forecasting function is able to forecast the true observation y_{161}. This we repeat by changing the forecasting origin to $T=161,\ldots,\ 167$.

For the purpose, we first attempt to fit Harvey and Fernandes (1989) random effects based dynamic model to this count series of length with first $T=160$ observations. As far as the time dependent covariates are concerned, we have used the same regression variables as in Zeger (1988). Consequently, we have regressed the monthly number of polio cases on a linear trend as well as sine, cosine pairs at annual and semi-annual frequencies to reveal the evidence of seasonality. More specifically, we use

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

where $t^{'}=(t-73)$ is used to locate the intercept term at January 1976, for $t=1,\cdots,160$. Note that the mean and variance of the 168 polio counts were found to be 1.33 and 3.48 respectively. This indicates the presence of overdispersion, and fitting the non-stationary negative binomial counts model appear to be appropriate.

In this approach, we attempt to maximize the log likelihood function (2.12) to estimate the regression parameter β and the overdispersion parameter w. Note that by assuming $a_0^*=b_0^*=0$, we first write $a_1=wa_0^*=0$ and $b_1=wb_0^*=0$. Next for $t=2,\ldots,T$, a_t and b_t may be expressed as

$$a_t=\sum_{j=1}^{t-1}w^j y_{t-j},\ \text{and}\ b_t=\sum_{j=1}^{t-1}w^j\mu_{t-j},$$

where $\mu_t=e^{x_t'\beta}$. By using these relationships in (2.12), we attempt to solve the log likelihood estimating equations for β and w, respectively, given by

$$\frac{\partial logL}{\partial\beta}=\sum_{t=2}^{T}\left[\{\frac{a_t}{b_t}\frac{\partial b_t}{\partial\beta}\}-\{\frac{a_t+y_t}{b_t+\mu_t}\}\frac{\partial(\mu_t+b_t)}{\partial\beta}\right]=0, \tag{2.16}$$

and

$$\frac{\partial logL}{\partial w}=\sum_{t=2}^{T}\left[\frac{\partial a_t}{\partial w}\{g_1(a_t+y_t)-g_1(a_t)+logb_t-log(b_t+\mu_t)\}+\frac{\partial b_t}{\partial w}\{\frac{a_t}{b_t}-\frac{a_t+y_t}{b_t+\mu_t}\}\right]=0, \tag{2.17}$$

with

$$\frac{\partial\mu_t}{\partial\beta}=\mu_t x_t,\ \frac{\partial b_t}{\partial\beta}=\sum_{j=1}^{t-1}w^j\mu_{t-j}x_{t-j},\ \frac{\partial a_t}{\partial w}=\sum_{j=1}^{t-1}jw^{j-1}y_{t-j},\ \text{and}\ \frac{\partial b_t}{\partial w}=\sum_{j=1}^{t-1}jw^{j-1}\mu_{t-j},$$

and, for example,

$$g_1(a_t) = \frac{\partial \Gamma(a_t)}{\partial a_t} \simeq log a_t - \frac{1}{2a_t} - \frac{1}{12a_t^2} + \frac{1}{120a_t^4}$$

(Abramowitz and Stegun (1965)).

Note however that starting with small initial values for the components of β vector as well as with a small value for w, we attempted to obtain the solutions of the above two equations (2.16) and (2.17) for β and w, but the equations did not yield any convergent solutions. This happened as this polio data appears to have a few large counts (see Figure 1) such as $y_t = 9,\ 14,\ 7,\ 8$ at time points $t = 7,\ 35,\ 113,\ 114$ respectively. Thus, this approach of Harvey and Farnendes (1989) does not appear to be a suitable approach to deal with other than low counts. This became evident from a re-analysis by replacing

[Insert Figure 1 about here]

these moderately large values by the mean 1 of the rest of the data. In this case, the log likelihood equations (2.16)-(2.17) yielded the estimates of the regression effects as

$$\hat{\beta}_1 = 0.25,\ \hat{\beta}_2 = -3.62,\ \hat{\beta}_3 = -0.01,\ \hat{\beta}_4 = -0.48,\ \hat{\beta}_5 = 0.20,\ \hat{\beta}_6 = -0.17$$

and the estimate for overdispersion parameter w as $\hat{w} = 0.90$. These likelihood estimates for the modified data appear to be close to the estimates found by Zeger (1988, col. 2, Table 3, p. 627) except for β_2 and β_6. When these estimates were used in the forecasting function (2.14) for one step ahead forecast we obtained forecasted values (after rounding to an integer)

$$\tilde{y}_{161} = 0,\ \tilde{y}_{162} = 1,\ \tilde{y}_{163} = 1,\ \tilde{y}_{164} = 1,\ \tilde{y}_{165} = 1,$$

whereas true counts were

$$y_{161} = 0,\ y_{162} = 1,\ y_{163} = 2,\ y_{164} = 1,\ y_{165} = 0,$$

respectively, showing that the forecasting function may work well for low counts. Note that as this approach appears to encounter difficulties with larger counts, in our simulation studies to be reported later we will not include this approach for comparison, rather, we will concentrate to the forecasting performance of the proposed model based approach for various sets of larger counts data.

2.3 Log-linear Dynamic Models

Note that as opposed to the parameter-driven models discussed in Sections 2.1 and 2.2, there exists an alternative log-linear dynamic model given by

$$f(y_t \mid y_1, \cdots, y_{t-1}) = \frac{e^{-\mu_t^*} \mu_t^{*y_t}}{y_t!}, \tag{2.18}$$

(Davis et al (2003)) where $\mu_t^* = \mu_t \theta_t$ with ARMA (r,q) model for logθ_t, i.e.,

$$log\theta_t = \gamma_t = \phi_1 \gamma_{t-1} + \cdots + \phi_r \gamma_{t-r} + e_t + \psi_1 e_{t-1} + \cdots + \psi_q e_{t-q}, \tag{2.19}$$

for example, where

$$e_t = \frac{y_t - \mu_t}{{\mu_t}^{-\lambda}}, \lambda \in (0, 1]. \tag{2.20}$$

Note however that as (2.19) is an auto-regression model for a sequence of random effects γ_t with errors defined based on the past observations, we prefer to call this model a log-linear dynamic (LLD) model, whereas Davis et al (2003) have referred to (2.18)-(2.20) as an observation-driven (OD) model. Further note that even though the LLD model (2.18)-(2.20) allows negative and positive autocorrelations $\rho_y(l)$ among count responses $y_1, \cdots, y_t, \cdots, y_T$, they are however neither easy to compute nor easy to interpret. The estimation of the parameters involved in this model (2.18)-(2.20) is also not easy.

2.3.1 Forecasting

Similarly to Zeger (1988), Davis et all (2003) also did not deal with the forecasting issues. Note that this LLD model is quite similar to that of the random effects based marginal model of Zeger (1988). The main difference between the two models is that in (2.1), the distribution of the random effects $\{\theta_t = e^{\gamma_t}\}$, whether multivariate gamma or multivariate log-normal , is independent of y_t $(t = 1, \ldots, T)$, whereas γ_t in Davis et all (2003) is defined as a function of the responses in their standardized form and $\{\theta_t = e^{\gamma_t}\}$ are assumed to have log-normal type distributions. Similar to Zeger, the computation for the forecasting function appears to be complicated. Hence we do not pursue this approach any further for the purpose of forecasting.

2.4 Observation-driven Stationary Poisson Correlation Models

For simplicity, in this sub-section and also in the next sub-section, we confine our discussion to the auto-regressive order 1 (AR(1)) case. The models for other correlation structures such as MA(1), equi-correlations, ARMA(1,1) may be developed similarly (Sutradhar (2003)).

Following Al-Osh and Alzaid (1987), and McKenzie (1988) [see also Freeland and McCabe (2004, p. 227-34)], one may write the Poisson auto-regressive dynamic model in the form

$$y_t = \rho * y_{t-1} + d_t, \tag{2.21}$$

where it is assumed that

$$y_{t-1} \sim P(\mu_.), d_t \sim P(\mu_.(1-\rho)) \tag{2.22}$$

and d_t and y_{t-1} are independent. In (2.22), $\mu_.$ is the constant Poisson mean parameter defined as $\mu_. = e^{x_.'\beta}$ which is time independent. Also in (2.21), for given count y_{t-1},

$$\rho * y_{t-1} = \sum_{j=1}^{y_{t-1}} b_j(\rho), \tag{2.23}$$

where $b_j(\rho)$ stands for a binary variable with $pr(b_j(\rho)=1)=\rho$ and $pr(b_j(\rho)=0)=1-\rho$. This operation in (2.21) is known as the so called binomial thinning operation. It can be shown that

$$y_t \sim P(\mu_.) \tag{2.24}$$

so that for the stationary data, i.e., for time independent covariates $x_t = x_.$ for all $t = 1, \ldots, T$, one writes

$$E(Y_t) = var(Y_t) = \mu_. = e^{x_.'\beta} \tag{2.25}$$

Furthermore, it can be shown that

$$\rho_y(l) = corr(Y_t, Y_{t-l}) = \rho^l, \tag{2.26}$$

where $0 < \rho < 1$.

2.4.1 Forecasting

It may be shown from (2.21)-(2.23) that for $\mu_\cdot = \exp(x'_\cdot\beta)$, the conditional distribution of y_t given y_{t-1} has the form given by

$$\begin{aligned} f_{t|t-1}(y_t|y_{t-1}) &= \exp\{-\mu_\cdot(1-\rho)\} \\ &\quad \times \sum_{k=0}^{\min(y_{t-1},y_t)} \frac{y_{t-1}!\rho^k {\mu_\cdot}^{y_t-k}(1-\rho)^{y_{t-1}+y_t-2k}}{k!(y_{t-1}-k)!(y_t-k)!}, \end{aligned} \tag{2.27}$$

see Freeland and McCabe (2004, Section 3, p. 428), also McKenzie (1988). Note that to develop the forecasting function, it is not necessary to derive the conditional distribution (2.27) . This is because by using (2.21) and (2.23) directly, the forecasting function can be written as

$$\begin{aligned} E(Y_{t+1}|y_t) &= E\left[\sum_{j=1}^{y_t} b_j(\rho)\right] + E(d_{t+1}) \\ &= \mu_\cdot + \rho(y_t - \mu_\cdot), \end{aligned} \tag{2.28}$$

with

$$\begin{aligned} \mathrm{var}(Y_{t+1}|y_t) &= \rho(1-\rho)y_t + (1-\rho)\mu_\cdot \\ &= \mu + \rho(y_t - \mu) - y_t\rho^2. \end{aligned} \tag{2.29}$$

Next, when the forecasting for lag 1 future count is made by (2.28), there occurs a forecasting error defined as

$$e_t(1) = y_{t+1} - E(Y_{t+1}|y_t), \tag{2.30}$$

which by (2.21), (2.23) and (2.28) may be expressed as

$$e_t(1) = \sum_{j=1}^{y_t} b_j(\rho) + d_{t+1} - [\mu_\cdot + \rho(y_t - \mu_\cdot)].$$

Consequently, the variance of the forecast error can be computed as

$$\begin{aligned} \mathrm{var}(e_t(1)) &= \mathrm{var}\left[E(e_t(1)|y_t)\right] + E\left[\mathrm{var}(e_t(1)|y_t)\right] \\ &= E\left[\mathrm{var}(Y_{t+1}|y_t)\right] \\ &= \mu_\cdot(1-\rho^2). \end{aligned} \tag{2.31}$$

2.5 Observation-driven Stationary Negative Binomial Correlation Models

To obtain correlated negative binomial count data, one may follow Lewis (1980) and McKenzie (1986) [see also Jowaheer and Sutradhar (2002)], and relate y_t with y_{t-1} by

$$y_t = \alpha_t * y_{t-1} + d_t, \tag{2.32}$$

where, for given probability $0 < \alpha_t < 1$ and count y_{t-1}, the symbol $*$ indicates the binomial thinning operation, so that $\alpha_t * y_{t-1}$ is the sum of y_{t-1} binary variables with probability α_t. That is,

$$z_t = \alpha_t * y_{t-1} = \sum_{j=1}^{y_{t-1}} b_j(\alpha_t), \tag{2.33}$$

where $b_j(\alpha_t)$ denotes the jth binary variable, with probability of success α_t, i.e. $\text{pr}\{b_j(\alpha_t) = 1\} = \alpha_t = 1 - \text{pr}\{b_j(\alpha_t) = 0\}$. It then follows that

$$[z_t|y_{t-1}, \alpha_t] \sim \text{Bi}(y_{t-1}, \alpha_t),$$

independently for all t. Next, under the assumption that $\alpha_t \sim \text{Be}\{\rho/c, (1-\rho)/c\}$, independently for all t, with $0 \leq \rho \leq 1$, that is,

$$f(\alpha_t) = \frac{\Gamma(1/c)}{\Gamma\{(1-\rho)/c\}\Gamma(\rho/c)} \int \alpha_t^{\rho/c-1}(1-\alpha_t)^{(1-\rho)/c-1} d\alpha_t,$$

one obtains the mean and the variance of α_t as

$$E[\alpha_t] = \rho, \text{ and } \text{var}[\alpha_t] = [\rho(1-\rho)c]/(1+c).$$

This yields the conditional mean and variance of z_t given y_{t-1} as

$$E(z_t|y_{t-1}) = E_{\alpha_t}(y_{t-1}\alpha_t) = \rho y_{t-1}, \tag{2.34}$$

and

$$\begin{aligned} \text{var}(z_t|y_{t-1}) &= E_{\alpha_t}[var(z_t|y_{t-1}, \alpha_t] + var_{\alpha_t}[E(z_t|y_{t-1}, \alpha_t)] \\ &= E_{\alpha_t}[\alpha_t(1-\alpha_t)y_{t-1}] + var_{\alpha_t}[\alpha_t y_{t-1}] \\ &= \rho(1-\rho)y_{i,t-1}\left[\frac{1+cy_{i,t-1}}{1+c}\right] \end{aligned} \tag{2.35}$$

Furthermore, suppose that in (2.32),

$$y_{t-1} \sim \text{NeBi}(1/c, c\mu_\cdot), \text{ and } d_t \sim \text{NeBi}\{(1-\rho)/c, c\mu_\cdot\}, \tag{2.36}$$

all variables being independent, with $\mu_\cdot = \exp(x_\cdot'\beta)$, where $x_\cdot$ is the $p \times 1$ vector of time independent covariates. Here

$$E[d_t] = (1-\rho)\mu_\cdot, \text{ and } \text{var}(d_t) = (1-\rho)\mu_\cdot[1 + c\mu_\cdot].$$

Now by applying (2.33) and (2.34), it follows from (2.32) that

$$E(y_t) = \mu_\cdot, \text{ and } var(y_t) = \mu_\cdot + c\mu_\cdot^2. \tag{2.37}$$

By similar calculations it follows from (2.32) that

$$E(y_t y_{t-\ell}) = \rho^\ell(\mu_\cdot + c\mu_\cdot^2) + \mu_\cdot^2, \tag{2.38}$$

yielding the lag ℓ auto-correlation as $\rho_\ell = \rho^\ell$.

2.5.1 Forecasting

Now by using the conditional expectation of z_{t+1} given y_t from (2.34) into (2.32), one obtains the conditional expectation of Y_{t+1} given y_t, and hence the forecasted value of y_{t+1} as

$$\begin{aligned}
\tilde{y}_{t+1} &= E[Y_{t+1}|y_t] \\
&= E_{\alpha_{t+1}} E[Y_{t+1}|y_t, \alpha_{t+1}] \\
&= E_{\alpha_{t+1}} [(1-\rho)\mu_\cdot + y_t\alpha_{t+1}] \\
&= (1-\rho)\mu_\cdot + \rho y_t \\
&= \mu_\cdot + \rho(y_t - \mu_\cdot),
\end{aligned} \tag{2.39}$$

where $\mu_\cdot = \mu_t$ for all $t = 1, \ldots, T$.

Note that as $E(\alpha_t) = \rho$, and $\text{var}(\alpha_t) = [\rho(1-\rho)c]/(1+c)$, the conditional variance of Y_{t+1} given y_t may be derived as

$$\begin{aligned}
var[Y_{t+1}|y_t] &= var_{\alpha_{t+1}} E[Y_{t+1}|y_t, \alpha_{t+1}] + E_{\alpha_{t+1}} var[Y_{t+1}|y_t, \alpha_{t+1}] \\
&= var_{\alpha_{t+1}} [(1-\rho)\mu_\cdot + y_t\alpha_{t+1}]
\end{aligned}$$

$$
\begin{aligned}
& + E_{\alpha_{t+1}}\left[y_t\alpha_{t+1}(1-\alpha_{t+1}) + (1-\rho)\{\mu_. + c\mu_.^2\}\right] \\
= \quad & \frac{\rho c(1-\rho)}{1+c} y_t^2 + E_{\alpha_{t+1}}\left[(1-\rho)\mu_.(1+c\mu_.) + y_t\alpha_{t+1}(1-\alpha_{t+1})\right] \\
= \quad & (1-\rho)\mu_.(1+c\mu_.) + \rho y_t(1-\rho)\left(\frac{1+cy_t}{1+c}\right) \qquad (2.40)
\end{aligned}
$$

Next similar to (2.30), the one-step ahead forecasting error is given by

$$e_t(1) = y_{t+1} - E(Y_{t+1}|y_t),$$

which by (2.32), (2.33) and (2.39) may be expressed as

$$e_t(1) = \sum_{j=1}^{y_t} b_j(\alpha_{t+1}) + d_{t+1} - [\mu_. + \rho(y_t - \mu_.)].$$

It then follows that

$$E[e_t(1)|y_t] = E_{\alpha_{t+1}} E[e_t(1)] = 0.$$

Consequently, the variance of the forecast error can be computed as

$$
\begin{aligned}
\mathrm{var}(e_t(1)) \quad = \quad & \mathrm{var}\left[E(e_t(1)|y_t)\right] + E\left[\mathrm{var}(e_t(1)|y_t)\right] \\
& = E\left[\mathrm{var}(Y_{t+1}|y_t)\right] \\
& = (1-\rho)[\mu_. + c\mu_.^2] + \rho(1-\rho)[\mu_. + c\mu_.^2] \\
& = (1-\rho^2)[\mu_. + c\mu_.^2] \qquad (2.41)
\end{aligned}
$$

which reduces to the variance of the forecasting error (2.31) under the Poisson model when $c \to 0$.

3 Proposed Observation-driven Non-stationary Correlation Models

In this section, we deal with the forecasting of a future count for non-stationary Poisson and negative binomial time series data. As far as the models for the non-stationary counts are concerned, we provide them in Section 3.1 for the Poisson data and in Section 3.2 for the negative binomial data. These models may be treated as a generalization of the stationary models considered by McKenzie

(1986, 1988), Al-Osh and Alzaid (1987), and Freeland and McCabe (2004), for example. For modelling non-stationary negative binomial time series data, we also refer to Mallick and Sutradhar (2004).

3.1 Observation-driven (OD) Non-stationary Poisson Models

Note that the stationary Poisson correlation model was discussed in Section 2.4. Recall from (2.21) that the Poisson auto-regressive dynamic model has the form given by

$$y_t = \rho * y_{t-1} + d_t. \tag{3.1}$$

In (3.1), we assume that

$$y_{t-1} \sim P(\mu_{t-1}), d_t \sim P(\mu_t - \rho\mu_{t-1}), \tag{3.2}$$

whereas in the stationary case it was assumed in (2.21)-(2.22) that

$$y_{t-1} \sim P(\mu_.), d_t \sim P((1-\rho)\mu_.).$$

In (3.1), d_t and y_{t-1} are independent. Now by using the binomial thinning operation (2.23), it follows from (3.1) and (3.2) that

$$y_t \sim P(\mu_t) \tag{3.3}$$

so that

$$E(Y_t) = var(Y_t) = \mu_t = e^{x_t^{'}\beta} \tag{3.4}$$

Furthermore, it can be shown that

$$\rho_y(l) = corr(Y_t, Y_{t-l}) = \rho^l \sqrt{\frac{\mu_{t-l}}{\mu_t}}. \tag{3.5}$$

Note that for the stationary case where $\mu_t = \mu_{t-l} = \mu_. = e^{x_.^{'}\beta}$ (say), the non-stationary correlation in (3.5) reduces to the stationary correlation given by $\rho_y(l) = \rho^l$ [McKenzie (1988, p.823-24), Sutradhar (2003, p.385-86)]. For the non-stationary case, it follows from (3.2) that for $\mu_t - \rho\mu_{t-1}$ to be non-negative ρ must satisfy the range $0 < \rho < \frac{\mu_t}{\mu_{t-1}}$. Consequently, even though in the stationary case ρ has the range $0 < \rho < 1$, in the non-stationary case, ρ must satisfy the range restriction

$$0 < \rho < min\left[1, \frac{\mu_t}{\mu_{t-1}}\right], t = 2, \cdots, T. \tag{3.6}$$

3.1.1 Forecasting

By using the model (3.1), we first write

$$y_{t+1} = \sum_{j=1}^{y_t} b_j(\rho) + d_{t+1}, \tag{3.7}$$

which for given y_t by (3.2) yields the forecasting function as

$$E(Y_{t+1}|y_t) = \mu_{t+1} + \rho(y_t - \mu_t). \tag{3.8}$$

Note that for given y_t, the conditional variance of the future observation y_{t+1} can easily be calculated, which has the formula given by

$$\begin{aligned} \text{var}(Y_{t+1}|y_t) &= \rho(1-\rho)y_t + [\mu_{t+1} - \rho\mu_t] \\ &= \mu_{t+1} + \rho(y_t - \mu_t) - \rho^2 y_t. \end{aligned} \tag{3.9}$$

Next it follows from (3.7)-(3.9) that the variance of the one step ahead forecasting error $e_t(1) = y_{t+1} - E(Y_{t+1}|y_t)$, has the formula given by

$$\begin{aligned} \text{var}(e_t(1)) &= \text{var}[E(e_t(1)|y_t] + E[\text{var}(e_t(1)|y_t)] \\ &= E[\text{var}(Y_{t+1}|y_t)] \\ &= \mu_{t+1} - \rho^2\mu_t. \end{aligned} \tag{3.10}$$

3.2 Observation-driven (OD) Non-stationary Negative Binomial Models

Recall from Section 2.5 that the negative binomial correlation model is given by

$$y_t = \alpha_t * y_{t-1} + d_t, \tag{3.11}$$

with $\alpha_t * y_{t-1} = \sum_{j=1}^{y_{t-1}} b_j(\alpha_t)$, where $b_j(\alpha_t)$ is a binary response with $Pr\left[b_j(\alpha_t) = 1\right] = \alpha_t$, where $\alpha_t \sim Be\{\rho/c, (1-\rho)/c\}$. Now unlike the stationary model discussed in Section 2.5, we assume that in (3.11), $y_{t-1} \sim NeBi(\frac{1}{c}, c\mu_{t-1})$ which reflects the non-stationarity of the count responses. As far as the distribution of d_t is concerned, we assume that

$$d_t \sim NeBi(\psi_1, \psi_2), \tag{3.12}$$

with

$$\psi_1 = \frac{(\mu_t - \rho\mu_{t-1})^2}{c(\mu_t^2 - \rho\mu_{t-1}^2)}, \text{ and } \psi_2 = \frac{c(\mu_t^2 - \rho\mu_{t-1}^2)}{(\mu_t - \rho\mu_{t-1})}$$

(Mallick and Sutradhar (2004)). Similar to (2.36), the first two moments of this distribution are given by

$$\begin{aligned} E[d_t] &= \psi_1\psi_2 = [\mu_t - \rho\mu_{t-1}] \\ \text{var}[d_t] &= \psi_1\psi_2(1+\psi_2) = [\mu_t - \rho\mu_{t-1}] + c[\mu_t^2 - \rho\mu_{t-1}^2] \end{aligned} \tag{3.13}$$

In general, by using the distribution of y_{t-1} given above, it can be shown that

$$y_t \sim NeBi\left(\frac{1}{c}, c\mu_t\right), \tag{3.14}$$

where $\mu_t = e^{x_t'\beta}$.

Furthermore, by using the relationship $y_t = \alpha_t * y_{t-1} + d_t$ (3.11), one may derive the correlation structure for non-stationary AR(1) type negative binomial counts as follows. As α_t's are independent with $E(\alpha_t) = \rho$, for all $t = 1, \cdots, T$, it then follows that $E(Y_tY_{t-1}) = \rho v_{t-1} + \mu_t\mu_{t-1}$, where $v_t = var(Y_t) = \mu_t + c\mu_t^2$, yielding lag-1 correlation $\rho_y(1) = \rho\sqrt{\frac{v_{t-1}}{v_t}}$. Also $E(Y_tY_{t-2}) = \rho^2 v_{t-2} + \mu_t\mu_{t-1}$, yielding lag-2 correlation $\rho_y(2) = \rho^2\sqrt{\frac{v_{t-2}}{v_t}}$. By similar calculation, one may show that, for $l = 1, \cdots, T-1$, the lag-l autocorrelation is given by

$$\rho_y(l) = \rho^l\sqrt{\frac{v_{t-l}}{v_t}}. \tag{3.15}$$

Thus the non-stationary negative binomial counts exhibit a non-stationary correlation structure which reduces to the Gaussian AR(1) type autocorrelation structure for the stationary negative binomial model. As far as the range restriction of ρ is concerned, it is clear that for ψ_1 and ψ_2 in (3.12) to be positive, ρ must satisfy

$$0 < \rho < min\left\{1, \frac{\mu_t}{\mu_{t-1}}, \frac{\mu_t^2}{\mu_{t-1}^2}\right\}, t = 2, \cdots, T. \tag{3.16}$$

Further, it is interesting to note that if $\mu_t = \mu.$ is used for all $t = 1, \cdots, T$, (3.11)-(3.12) yields the distributions of y_{t-1} and d_t as $y_{t-1} \sim NeBi\left(\frac{1}{c}, c\mu.\right)$, $d_t \sim NeBi\left(\frac{1-\rho}{c}, c\mu.\right)$. These special distributional assumptions were used in McKenzie (1986) and Jowaheer and Sutradhar (2002) to derive the correlation model for stationary negative binomial data.

3.2.1 Forecasting

By using the model (3.11), we first write

$$y_{t+1} = \sum_{j=1}^{y_t} b_j(\alpha_t) + d_{t+1}, \tag{3.17}$$

which for given y_t yields the forecasting function as

$$\begin{aligned} E(Y_{t+1}|y_t) &= E_{\alpha_t}[y_t\alpha_t] + [\mu_{t+1} - \rho\mu_t] \\ &= \mu_{t+1} + \rho(y_t - \mu_t), \end{aligned} \tag{3.18}$$

as $E(\alpha_t) = \rho$.

Note that the forecasting function in (3.17) is the same as the forecasting function (3.8) under the non-stationary Poisson model. Further note that as $\text{var}(\alpha_t) = [\rho(1-\rho)c]/(1+c)$, for given y_t, one may compute the conditional variance of the future observation y_{t+1} as

$$\begin{aligned} \text{var}(Y_{t+1}|y_t) &= E_{\alpha_t}[\text{var}(Y_{t+1}|y_t, \alpha_t)] + \text{var}_{\alpha_t}[E(Y_{t+1}|y_t, \alpha_t)] \\ &= E_{\alpha_t}[y_t\alpha_t(1-\alpha_t) + (\mu_{t+1} - \rho\mu_t) + c(\mu_{t+1}^2 - \rho\mu_t^2)] \\ &\quad + \text{var}_{\alpha_t}[y_t\alpha_t + \mu_{t+1} - \rho\mu_t] \\ &= \frac{\rho(1-\rho)}{1+c} y_t[1+cy_t] + [(\mu_{t+1} - \rho\mu_t) + c(\mu_{t+1}^2 - \rho\mu_t^2)] \end{aligned} \tag{3.19}$$

This conditional variance reduces to (2.40) for the stationary case where $\mu_t = \mu.$ for all $t = 1, \ldots, T$. Next by (3.17)-(3.18) we compute the variance of the one step ahead forecasting error $e_t(1) = y_{t+1} - E(Y_{t+1}|y_t)$. The formula for this variance is given by

$$\begin{aligned} \text{var}(e_t(1)) &= \text{var}[E(e_t(1)|y_t] + E[\text{var}(e_t(1)|y_t)] \\ &= E[\text{var}(Y_{t+1}|y_t)] \\ &= \rho(1-\rho)[\mu_t + c\mu_t^2] + [(\mu_{t+1} - \rho\mu_t) + c(\mu_{t+1}^2 - \rho\mu_t^2)]. \end{aligned} \tag{3.20}$$

4 Estimation of Parameters

As the forecasting functions contain unknown parameters of the model, in this section, we propose a generalized quasilikelihood (GQL) approach for the esti-

mation of the parameters of both non-stationary Poisson and negative binomial mixed models. This approach yields consistent estimators for the respective parameters. Note that we do not provide any estimating formulas for the stationary models as they are special cases of the respective non-stationary models.

4.1 GQL Estimation Approach For Non-stationary Poisson Mixed Model

The GQL approach exploits the mean vector and the covariance structure of the data. To be specific, let $y = (y_1, \cdots, y_t, \cdots, y_T)^{'}$ be the T-dimensional vector of all responses that follow the non-stationary Poisson mixed model (NSPMM) discussed in Section 3.1 . Under this model, the marginal means, variances and lag ℓ correlations are given by (3.4) and (3.5). Let $\mu = (\mu_1, \cdots, \mu_t, \cdots, \mu_T)^{'}$ be the mean vector of y, where by (3.4), $\mu_t = e^{x_t^{'}\beta}$. Furthermore, let $\Sigma = (\sigma_{tt'})$ be the $T \times T$ covariance matrix of y, where

$$\sigma_{tt'} = \begin{cases} \mu_t, & \text{if } t = t^{'} \\ \rho^{|t-t^{'}|}\mu_t, & \text{if } t < t^{'} \end{cases} . \tag{4.1}$$

It then follows that for known ρ, one may write the GQL estimating equation (Sutradhar (2003, Section 3.1)) for β as

$$\frac{\partial \mu^{'}}{\partial \beta}\Sigma^{-1}(y - \mu) = 0, \tag{4.2}$$

which may be solved iteratively by Newton-Raphson iterative technique. To be specific, (4.2) is solved for β iteratively by using

$$\hat{\beta}(r+1) = \hat{\beta}(r) + \left[\left(X^{'}A\Sigma^{-1}AX\right)^{-1} X^{'}A\Sigma^{-1}(y-\mu)\right]_{[r]}, \tag{4.3}$$

where $A = diag(\mu_1, \cdots, \mu_t, \cdots, \mu_T)$, $X = (x_1^{'}, \cdots, x_t^{'}, \cdots, x_T^{'})^{'}$ with $x_t = (x_{t1}, x_{t2}, \cdots, x_{tp})^{'}$, and $[.]_r$ denotes the fact that the expression within the brackets is evaluated at $\hat{\beta}(r)$. Let $\hat{\beta}_{GQL}$ denote the solution obtained from (4.3).

Note that for the estimation of β by (4.3), it was assumed that the ρ parameter is known. As this stationary correlation parameter is however unknown in practice, we may estimate this parameter consistently by using the well known moment method. For this, we first observe that $E(Y_t - \mu_t)(Y_{t-1} - \mu_{t-1}) = \rho\mu_{t-1}$ for $t = 2, \ldots, T$. Consequently, the stationary lag-1 correlation parameter ρ may

be estimated by equating the lag-1 correlation with its sample counterpart. This provides the moment equation of ρ as

$$\hat{\rho} = \frac{\sum_{t=2}^{T} \tilde{y}_t \tilde{y}_{t-1}}{\sum_{t=1}^{T} \tilde{y}_t^2} \frac{T}{\sum_{t=2}^{T} \left[\mu_{t-1}/\mu_t\right]^{1/2}}, \tag{4.4}$$

where $\tilde{y}_t = \frac{y_t - \mu_t}{\sqrt{v_t}}$ with $v_t = \text{var}(y_t) = \sigma_{tt} = e^{(x_t'\beta)}$.

Note $\hat{\rho}$ obtained by (4.4) is consistent as it is obtained from an unbiased moment estimating equation. The performance of the GQL estimation approach is examined through a simulation study in Section 6. The performance of the forecasting function (3.8) is also examined by a simulation study in the same section. The forecasting approach based on GQL estimation methodology is illustrated in Section 5 by re-analyzing the US polio data that was analyzed earlier by Zeger (1988) and Davis et al (2000).

4.2 GQL Approach For the Estimation of the Parameters of Non-stationary Negative Binomial Mixed Model

Under the non-stationary negative binomial mixed model (NSNBMM) discussed in Section 3.2, the means and variances are given by $\mu_t = e^{(x_t'\beta)}$ and $v_t = \text{var}(y_t) = \mu_t + c\mu_t^2$ respectively. The formula for lag ℓ correlation is given by (3.15).

Similar to the NSPMM, we write the GQL estimating equation for β as

$$\frac{\partial \mu'}{\partial \beta} \Sigma^{-1} (y - \mu) = 0, \tag{4.5}$$

where μ is the same vector as in the Poisson case, but Σ matrix is now given by

$$\sigma_{tt'} = \begin{cases} v_t, & \text{if } t = t' \\ \rho^{|t-t'|} v_t, & \text{if } t < t' \end{cases} . \tag{4.6}$$

where $v_t = \mu_t + c\mu_t^2$.

Note that the estimating equation (4.5) requires the over-dispersion parameter c and the stationary correlation parameter ρ to be known. These parameters are however unknown in practice, which may be consistently estimated by using the well-known method of moments. To be specific, as $E(Y_t - \mu_t)^2 = v_t = \mu_t + \alpha\mu_t^2$, with $\mu_t = e^{x_t'\beta}$, one obtains the moment equation of c as

$$\hat{c} = \frac{\sum_{t=1}^{T} \left[(y_t - \hat{\mu}_t)^2 - \hat{\mu}_t\right]}{\sum_{t=1}^{T} \hat{\mu}_t^2}, \tag{4.7}$$

where $\hat{\mu}_t = e^{x_t^{'}\hat{\beta}_{GQL}}$.

As far as the moment estimation of ρ is concerned, we first observe that $E(Y_t - \mu_t)(Y_{t-1} - \mu_{t-1}) = \rho[\mu_{t-1} + c\mu_{t-1}^2]$. Consequently, the stationary lag-1 correlation parameter ρ may be estimated by equating the lag-1 correlation with its sample counterpart. This provides the moment equation of ρ as

$$\hat{\rho} = \frac{\sum_{t=2}^{T} \tilde{y}_t \tilde{y}_{t-1}}{\sum_{t=1}^{T} \tilde{y}_t^2} \frac{T}{\sum_{t=2}^{T} [v_{t-1}/v_t]^{1/2}}, \tag{4.8}$$

where $\tilde{y}_t = \frac{y_t - \mu_t}{\sqrt{v_t}}$, with $v_t = \mu_t + c\mu_t^2$.

Note that both $\hat{c}$ and $\hat{\rho}$ obtained by (4.7) and (4.8) respectively are consistent as they are obtained from unbiased estimating equations. Further note that under the present NSNBMM, $\hat{\rho}$ must satisfy the range restriction

$$0 < \hat{\rho} < \min\left[1, \frac{\hat{\mu}_t}{\hat{\mu}_{t-1}}, \frac{\hat{\mu}_t^2}{\hat{\mu}_{t-1}^2}\right], t = 2, \ldots, T.$$

5 Forecasting Polio Counts By Using OD Non-stationary Poisson and Negative Binomial Models

Recall from Section 2.2.3 that after replacing several large counts by the mean of the other observations, the dynamic model due to Harvey and Fernandes (1989) were applied to make one step ahead forecast for the future polio counts. This replacement however appears to be arbitrary. In this section, we first fit the observation driven non-stationary Poisson and negative binomial models to the U.S. polio count data with first $T = 160$ unmodified/original observations. These models are introduced in Sections 3.1 and 3.2, respectively.

Starting with small initial values for the regression and correlation parameters, we applied the GQL iterative procedure described in Section 4.1 for under the Poisson model. More specifically, having chosen starting values of zero for correlation and small values for regression parameters, we used (4.3) to obtain a convergent estimate of β. We then used this first step β estimate in (4.4) for ρ. Note that these first step estimates are in fact 1-cycle based estimates.

[Insert Table 1 about here]

We have continued this cycle of iterations until convergence for estimates of all parameters. It was found that the convergence was achieved in 5 cycles of iterations. These converged results are shown in column 2 of Table 1. The standard errors of the estimates under the Poisson model are given in column 3 of the same table.

We have also fitted the negative binomial correlation model to this polio counts data set. More specifically, similar to the Poisson case, the regression, overdispersion, and the correlation parameters were iteratively obtained by solving (4.5), (4.7) and (4.8). The convergent estimates along with their standard errors are given in columns 6-7 in Table 1. The regression and the correlation estimates under the two models appear to be close to each other, the standard errors are being slightly larger under the negative binomial model, as expected.

Next, to examine the performances of the forecasting function (3.8) under the correlated Poisson model and of (3.18) under the correlated negative binomial model, we have computed these functions by using the estimates of β and ρ under the respective models. For forecasting origin $T = 160$, 161, 162, 163, and 164 the one step ahead forecasted (OSAF) values along with corresponding true values are shown in columns 5 and 4 under the Poisson model and in columns 9 and 8 under the negative binomial model. It is clear that the forecasting functions under these two models yielded the same forecasted values. In fact, these forecasts are same as those produced by the dynamic model used by Harvey and Fernandes (1989) (see Section 2.2.3). Nevertheless, we recommend the use of the proposed observation driven (OD) Poisson and negative binomial correlation models based forecasting function in forecasting the future counts. This is mainly because, unlike the dynamic model, the OD correlation models do not appear to encounter any problems in fitting low as well as larger counts. In the next section, we verify this observation through a simulation study by forecasting one step ahead count for various count pattern data.

6 Observation Driven Correlation Models Based Forecasting : A Simulation Study

As compared to the existing models due to Zeger (1988), Davis et all (2003), Harvey and Fernandes (1989) (see also Settimi and Smith (2000)), the proposed observation driven (OD) Poisson and negative binomial models (see also

Freeland and McCabe (2004)) are discrete analogues of the well known Gaussian auto-regressive models, for fitting the time series of counts. The existing dynamic model (DM) due to Harvey and Fernandes (1989) as well as the proposed OD models were fitted to the U.S. polio count data before making any forecasts. As opposed to the OD models, the DM appear to have serious problems in fitting time series with moderately large or larger counts. Now to make sure about the fitting and forecasting performances of the proposed OD models based approaches, in this section, we conduct a simulation study involving time series of counts with low, moderately large and large counts.

To generate a time series of counts of length $T = 100$ under non-stationary Poisson model (3.1) and negative binomial model (3.11), we have first chosen a non-stationary regression model with $\mu_t = e^{(x_t'\beta)}$, where for simplicity, a two dimensional ($p = 2$) $\beta = (\beta_1,\ \beta_2)'$ is considered along with $x_t' = (x_{t1}, x_{t2})$ with $x_{t1} = 1$ for all $t = 1, \ldots, T$ but x_{t2} was chosen to be time dependent as

$$x_{t2} = \begin{cases} 0.01 & \text{for } t = 1 \\ x_{t-1,2} + 0.01 & \text{for } t = 1, \ldots, T/4 \\ x_{t-1,2} + 0.05 & \text{for } t = T/4 + 1, \ldots, 3T/4 \\ x_{t-1,2} + 0.10 & \text{for } t = 3T/4 + 1, \ldots, T \end{cases}$$

For the simulations under the Poisson model, we have chosen three sets of parameter values: (1) $\beta_1 = 0.5,\ \beta_2 = 0.5$, (2) $\beta_1 = -0.5,\ \beta_2 = 0.5$, and (3) $\beta_1 = 0.5,\ \beta_2 = -0.5$. The first set of parameter values generated large counts with non-stationary Poisson means ranging from 1.66 to 23.93. Similarly, the second and the third sets of parameter values generated Poisson counts with means ranging from 0.61 to 8.80, and 1.64 to 0.11, respectively. Note that for the third set, counts were generated in decreasing order. In each case, we have considered two values of correlation parameter, namely $\rho = 0.5$ and 0.8. Under the negative binomial model, we have chosen two sets of parameter values, namely, (1) $\beta_1 = 0.0,\ \beta_2 = 0.1$, and (2) $\beta_1 = 0.0,\ \beta_2 = 0.4$. These values along with the chosen covariate values generated counts with means ranging from 1.0 to 1.7 and 1.o to 8.5, respectively. As far as the overdispersion parameter is concerned, we have generated data with three selected values of $c = 0.20,\ 0.50,\ 1.00$. With regard to the correlation parameter, we have chosen $\rho = 0.5$ and $\rho = 0.9$. To have a feel about the

[Insert Figures 2 and 3 about here]

magnitude of the generated counts, we have displayed the true non-stationary mean values in Figure 2 for the Poisson case with first set of regression parameter values, and in Figure 3 for the negative binomial case with second set of parameter values.

The GQL estimates of the parameters were obtained following the formulas given in Section 4. We have conducted 500 simulations. The simulated means and standard errors for the estimates of each parameter value are given in Table 2 for the Poisson model and in Table 3 for the negative binomial model.

[Insert Tables 2 and 3 about here]

It is clear from the Table 2 that the GQL approach performs extremely well in estimating both β and ρ parameters under the Poisson mixed model. The GQL approach also estimate these parameters very well (see Table 3) under the negative binomial model. As far as the estimation of the overdispersion parameter c for negative binomial model is concerned, the moment approach as a part of the GQL approach appears to underestimate this parameter. A clustered GQL approach (see Mallick and Sutradhar (2004)) may be used to improve this estimate but we have not included this approach in this report for simplicity.

Next to examine the forecasting performances of the forecasting functions (3.8) for the Poisson model and of (3.18) for the negative binomial model, we have generated 101 counts but fitted the models to the first $T = 100$ counts and forecasted the one step ahead forecast at time point $T = 101$, under each simulation. The simulated means, standard errors of the true counts y_{101}, forecasted counts $\tilde{y}_{101}$, and of the corresponding forecasting errors, are shown in columns 7-9 in Table 2 under the Poisson model and in columns 9-11 in Table 3 under the negative binomial model. It is clear from the tables that the simulated means of the true counts are quite close to the simulated means of the forecasted counts indicating that the proposed forecasting functions works very well in forecasting a future count. When the counts are large, the forecasting errors appear to have larger standard errors as compared to the cases with time series of small counts, as expected. To have a feel about the differences between true counts and their forecasted counterparts, we have displayed the simulated counts in Figure 4 for 50 simulations under the Poisson model and in Figure 5 under the negative binomial model. The differences between the true and forecasted counts appear to be negligible.

[Insert Figures 4 and 5 about here]

7 Conclusion

As opposed to the random effects based parameters driven models of Zeger (1988) and Davis et al (2000, 2003), and the random effects based dynamic model of Harvey and Fernandes (1989)[see also Settimi and Smith (2000)], in this report, we have introduced a simple observation driven correlation models for both non-stationary time series of Poisson and negative binomial counts. The proposed model may be treated as a generalization of the stationary model considered by Freeland and McCabe (2004). The performances of the proposed one step ahead forecasting functions are examined through a simulation study and it is shown that they perform quite well. The proposed forecasting methodology is also illustrated by re-analyzing the time series of U.S. polio counts, earlier analyzed by Zeger (1988) and Davis et al (2000).

Acknowledgements

The author gratefully acknowledges the reserach grant for this work provided by SAS and the IIF. This research was also partially supported by a grant from Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Abramowitz, M. and Stegun, I. A. (Eds.) (1965). *Handbook of Mathematical Functions.* Dover, New York.

Al-osh, M. A. and Aly, E. A. A. (1992). First order autoregressive time series with negative binomial and geometric marginals. *Commun. Statist.-Theory Meth.*, **21**, 2483-2492.

Cox, D.R. (1970). *Analysis of Binary Data.* London: Chapman and Hall.

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

Davis, R.A., Dunsmuir, W.T.M. and Streett, S.B. (2003). Observation-driven models for Poisson counts. *Biometrika*, **90**, 777-790.

Freeland, R. K. and McCabe, B. P. M. (2004). Forecasting discrete valued low count time series. *International Journal of Forecasting*, **20**, 427-434.

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

Jacobs, P. A. and Lewis, P. A. W. (1978a). Discrete time series generated by mixtures I: correlational and run properties. *J. R. Stat. Soc., Ser. B*, **40**, 94-105.

Jacobs, P. A. and Lewis, P. A. W. (1978b). Discrete time series generated by mixtures II: asymptotic properties. *J. R. Stat. Soc., Ser. B*, **40**, 222-228.

Jowaheer, V. and Sutradhar, B. C. (2002). Analysing longitudinal count data with overdispersion. *Biometrika*, **89**, 389-399.

Kanter, M. (1975). Autoregression for discrete processes mod 2. *J. Appl. Probab.*, **12**, 371-375.

Kedem, B. (1980). Estimation of the parameters in stationary autoregressive process after hard limiting. *J. Am. Stat. Assoc.*, **75**, 146-153.

Keenan, D. M. (1982). A time series analysis of binary data. *J. Am. Stat. Assoc.*, **77**, 816-821.

Kulendran, N. and King, M. L. (1997). Forecasting international quarterly tourist flows using error correction and time series models. *International Journal of Forecasting*, **13**, 319-327.

Mallick, T. and Sutradhar, B. C. (2004). Analyzing time series of non-stationary negative binomial counts. *Technical Report, 8*, Department of Mathematics and Statistics, memorial University of Newfoundland, Canada.

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

McKenzie, E. (1988). Some ARMA models for dependent sequences of Poisson counts. *Adv. in Appl. Probab.*, **20**, 822-835.

Sutradhar, B.C. (2003). An overview on regression models for discrete longitudinal responses. *Statistical Science*, **18**, 377-393.

Sim, C. H. and Lee, P. A. (1989). Simulation of negative binomial processes. *J. Statist. Comput. Simul.*, **34**, 29-42.

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

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

Harvey, A. C. and Fernandes, G. (1989), "Time Series Models for Count or Qualitative Observations," *Journal of Business and Economics Statistics*, 7, 407-417.

Jowaheer, V. and Sutradhar, B. C. (2002), "Analysing Longitudinal Data with Overdispersion," *Biometrika*, 89, 389-399.

Kulendran, N. and King, M. L. (1997), "Forecasting International Quarterly Tourist Flows Using Error-Correction Models and Time-Series Models," *International Journal of Forecasting*, 13, 319-327.

McKenzie, E. (1986), "Autoregressive Moving-average Processes with Negative Binomial and Geometric Marginal Distributions," *Advances in Applied Probability*, 18, 679-705.

Settimi, R. and Smith, J. Q. (2000), "A Comparison of Approximate Bayesian Forecasting Methods for Non-Gaussian Time Series," *Journal of Forecasting*," 19, 135-148.

Table 1: Proposed observation-driven Poisson and negative binomial correlation models based GQL parameter estimates and one step ahead forecasted values at time point T for U.S polio count data (Zeger, 1988).

	Poisson Model				Negative Binomial Model			
	GQLE		OSAF		GQLE		OSAF	
Parameters	EST	SE	TRUE	FORECAST	EST	SE	TRUE	FORECAST
Intercept (β_1)	0.19	0.09	–	–	0.19	0.13	–	–
Trend$\times 10^{-3}$ (β_2)	-5.89	1.94	–	–	-5.02	2.83	–	–
$\cos(2\pi t/12)(\beta_3)$	-0.19	0.12	–	–	-0.19	0.18	–	–
$\sin(2\pi t/12)(\beta_4)$	-0.51	0.13	–	–	-0.46	0.19	–	–
$\cos(2\pi t/6)(\beta_5)$	0.12	0.11	–	–	0.11	0.16	–	–
$\sin(2\pi t/6)(\beta_6)$	-0.40	0.11	–	–	-0.37	0.16	–	–
c	–	–	–	–	0.85	–	–	–
ρ	0.23	–	–	–	0.22	–	–	–
T=161	–	–	0	1	–	–	0	1
T=162	–	–	1	1	–	–	1	1
T=163	–	–	2	1	–	–	2	1
T=164	–	–	1	1	–	–	1	1
T=165	–	–	0	0	–	–	0	0

Table 2: Simulated GQL estimates and one-step ahead forecasts for Poisson counts with both monotonic increasing and decreasing mean patterns for $T = 100$ and selected values of longitudinal correlation parameter, based on 500 simulations.

			Simulated estimates and one-step ahead forecast					
Regression Parameters	ρ		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\rho}$	y_{101}	$\tilde{y}_{101}$	$\hat{e}_{100}(1)$
$\beta_1 = 0.5,\ \beta_2 = 0.5$	0.50	SM	0.50	0.50	0.47	23.82	23.94	-0.12
$\mu_1 = 1.66, \ldots, \mu_{101} = 23.93$		SSE	0.16	0.04	0.10	4.72	3.28	4.45
	0.80	SM	0.45	0.51	0.75	23.98	24.16	-0.17
		SSE	0.24	0.07	0.07	5.35	4.85	3.33
$\beta_1 = -0.5,\ \beta_2 = 0.5$	0.50	SM	-0.52	0.50	0.45	8.76	8.75	0.01
$\mu_1 = 0.61, \ldots, \mu_{101} = 8.80$		SSE	0.27	0.07	0.10	2.89	2.05	2.70
	0.80	SM	-0.56	0.51	0.75	8.62	8.71	-0.09
		SSE	0.44	0.11	0.08	2.88	2.61	1.87
$\beta_1 = 0.5,\ \beta_2 = -0.5$	0.50	SM	0.48	-0.54	0.44	0.15	0.13	0.02
$\mu_1 = 1.64, \ldots, \mu_{101} = 0.11$		SSE	0.24	0.20	0.11	0.40	0.19	0.34
	0.80	SM	0.50	-0.62	0.72	0.11	0.12	-0.01
		SSE	0.42	0.40	0.13	0.33	0.28	0.22

Table 3: Simulated GQL estimates and one-step ahead forecasts for negative binomial counts with two monotonic increasing mean patterns for $T = 100$ and selected values of overdispersion and longitudinal correlation parameters, based on 500 simulations.

				Simulated estimates and one-step ahead forecast						
Regression Parameters	ρ	c		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\rho}$	$\hat{c}$	y_{101}	$\tilde{y}_{101}$	$\hat{e}_{100}(1)$
$\beta_1 = 0.0,\ \beta_2 = 0.1$	0.50	0.20	SM	-0.03	0.10	0.44	0.14	1.67	1.79	-0.12
$\mu_1 = 1.0, \cdot, \mu_{101} = 1.7$			SSE	0.29	0.11	0.11	0.14	1.53	0.92	1.41
		0.50	SM	-0.04	0.10	0.44	0.41	1.76	1.75	0.01
			SSE	0.32	0.13	0.12	0.27	1.75	1.04	1.69
		1.00	SM	-0.03	0.09	0.43	0.85	1.71	1.76	-0.05
			SSE	0.36	0.14	0.13	0.48	2.09	1.29	1.92
	0.75	0.20	SM	-0.06	0.09	0.68	0.16	1.60	1.69	-0.08
			SSE	0.41	0.16	0.09	0.14	1.38	1.23	1.07
		0.50	SM	-0.05	0.09	0.68	0.40	1.83	1.87	-0.04
			SSE	0.46	0.18	0.09	0.34	1.84	1.62	1.20
		1.00	SM	-0.05	0.09	0.67	0.81	1.94	1.96	-0.04
			SSE	0.56	0.24	0.11	0.59	2.31	1.92	1.78
$\beta_1 = 0.0,\ \beta_2 = 0.4$	0.50	0.20	SM	-0.04	0.41	0.45	0.16	8.56	8.65	-0.07
$\mu_1 = 1.0, \cdot, \mu_{101} = 8.5$			SSE	0.26	0.08	0.10	0.11	4.87	2.98	4.83
		0.50	SM	-0.04	0.40	0.43	0.41	8.41	8.68	-0.26
			SSE	0.29	0.10	0.11	0.21	6.36	4.17	6.38
		1.00	SM	-0.05	0.40	0.44	0.83	7.56	8.65	-1.09
			SSE	0.35	0.13	0.13	0.39	7.39	5.25	7.41
	0.75	0.20	SM	-0.04	0.39	0.69	0.13	8.49	8.44	0.05
			SSE	0.37	0.12	0.09	0.12	4.97	4.00	3.80
		0.50	SM	-0.07	0.41	0.68	0.38	8.63	8.75	-0.12
			SSE	0.43	0.14	0.10	0.23	6.23	4.97	4.40
		1.00	SM	-0.07	0.38	0.67	0.80	8.15	8.54	-0.39
			SSE	0.49	0.18	0.12	0.46	8.53	7.29	5.72

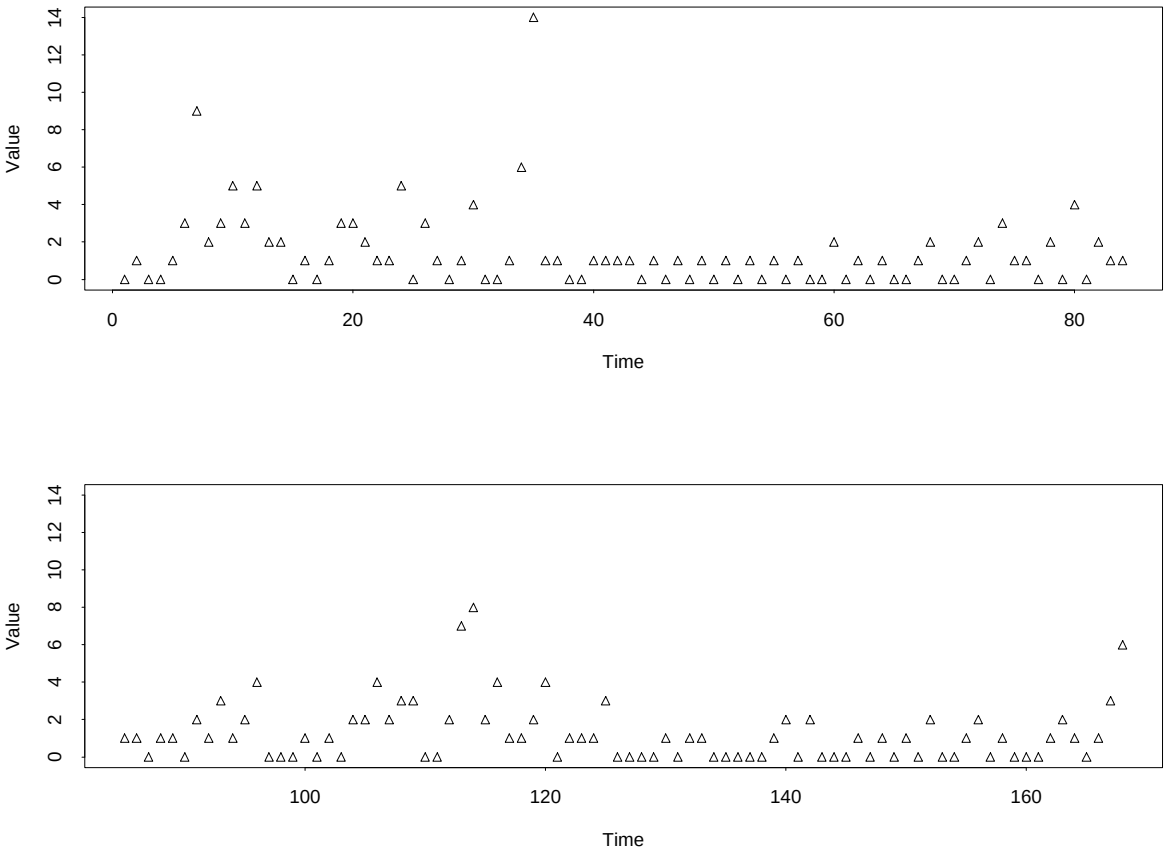

Figure 1: U.S. polio count data from January 1970 to December 1983.

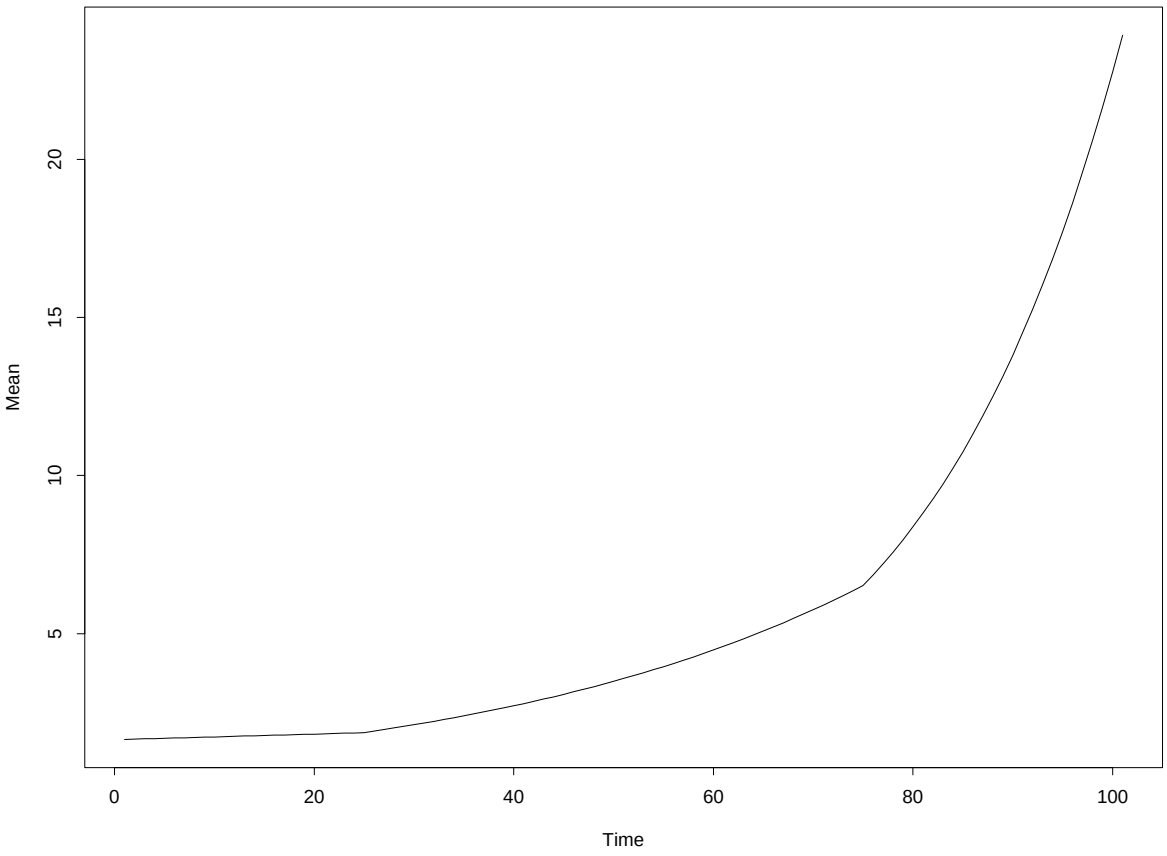

Figure 2: Non-stationary Poisson mean pattern with $\beta_1 = \beta_2 = 0.5$.

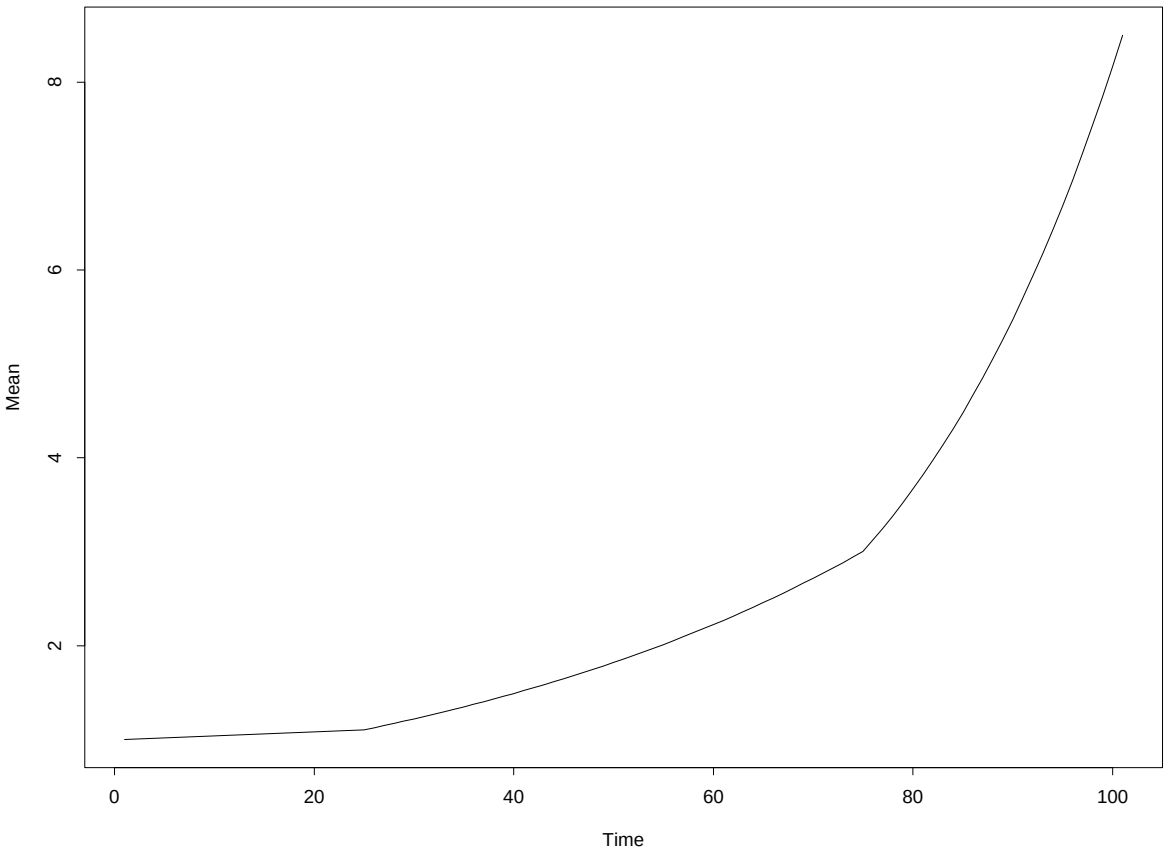

Figure 3: Non-stationary negative binomial mean pattern with $\beta_1 = 0$, $\beta_2 = 0.4$.

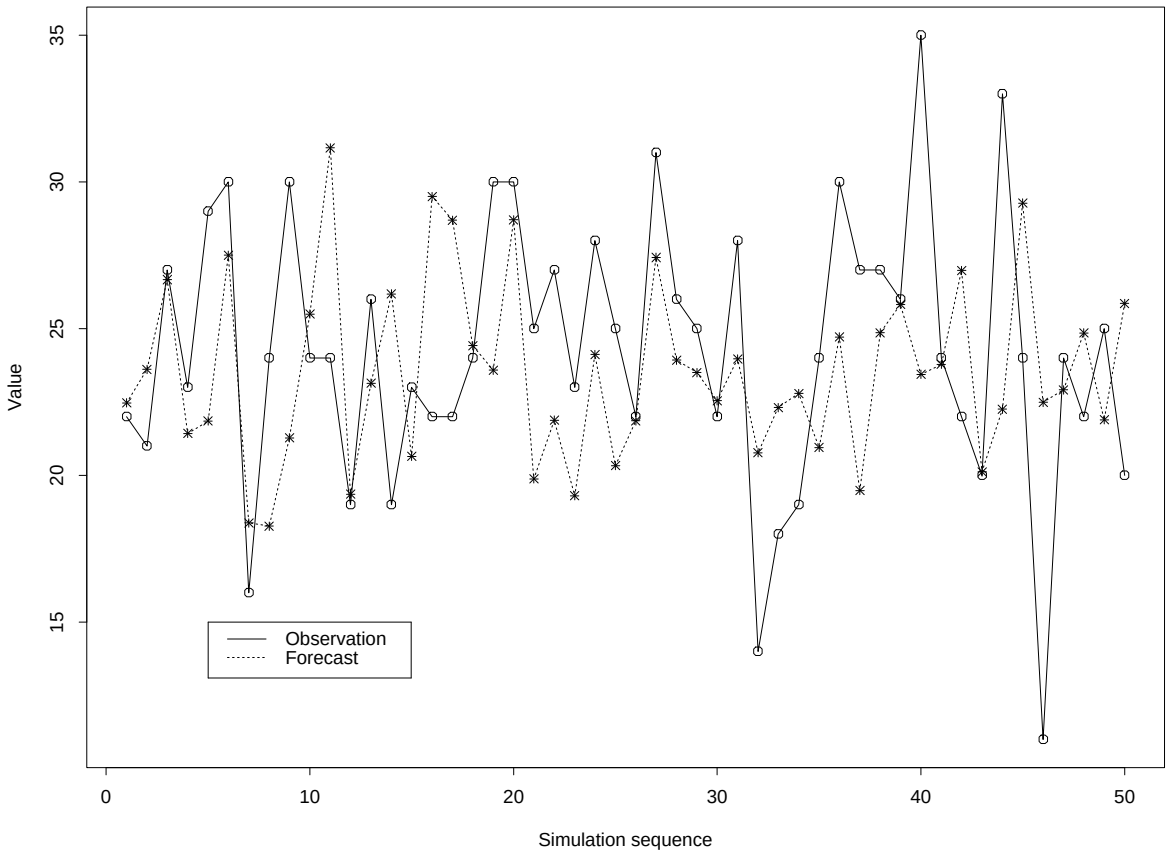

Figure 4: Simulated true and forecasted Poisson counts.

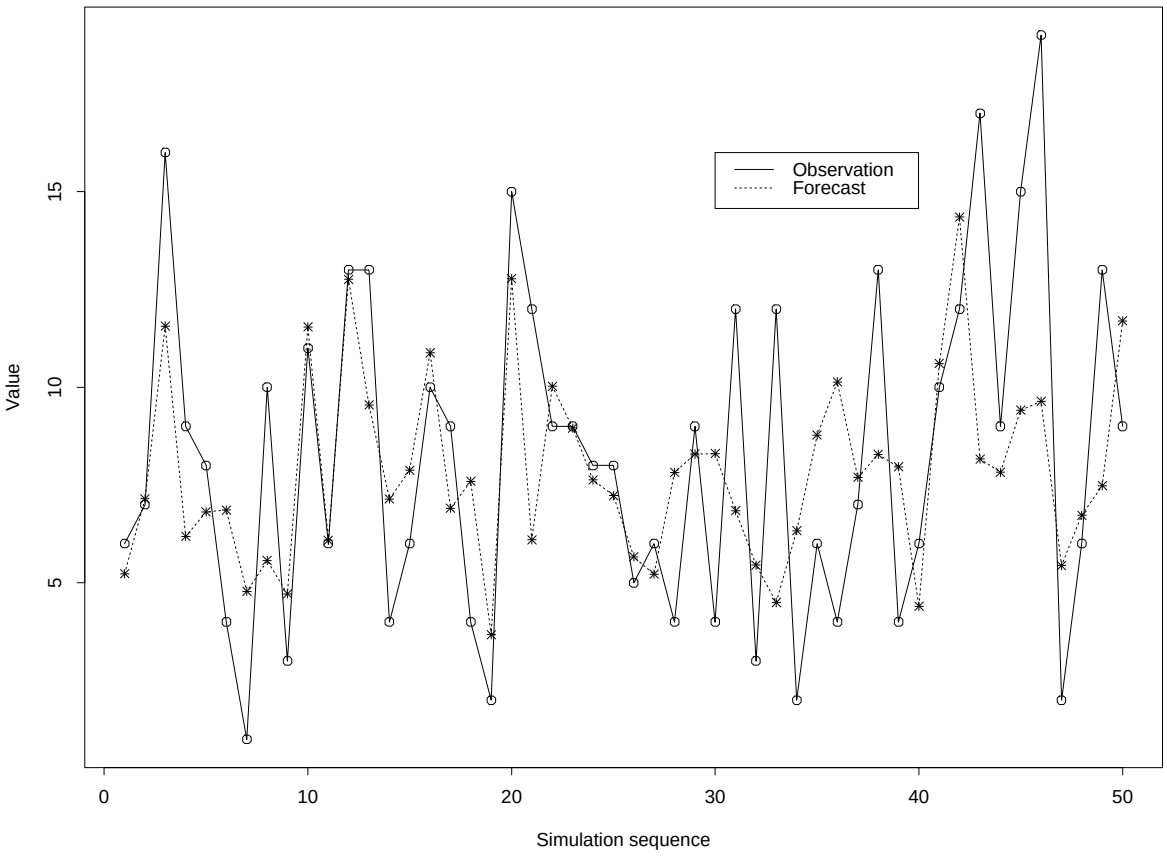

Figure 5: Simulated true and forecasted negative binomial counts.