Journal of Forecasting, Vol. 16, 195–207 (1997)

Predictions in Overdispersed Series of Counts Using an Approximate Predictive Likelihood

PHILIPPE LAMBERT*
University of Liège, Belgium

ABSTRACT

The generalized autoregression model or GARM, originally used to model series of non-negative data measured at irregularly spaced time points (Lambert, 1996a), is considered in a count data context. It is first shown how the GARM can be expressed as a GLM in the special case of a linear model for some transform of the location parameter. The Butler approximate predictive likelihood (Butler, 1986, Rejoinder) is then used to define likelihood prediction envelopes. The width of these intervals is shown to be slightly wider than the Fisher (1959, pp. 128–33) and Lejeune and Faulkenberry (1982) predictive likelihood-based envelopes which assume that the parameters have fixed known values (equal to their maximum likelihood estimates). The method is illustrated on a small count data set showing overdispersion. © 1997 by John Wiley and Sons, Ltd.

J. forecast. **16**: 195–207, 1997

No. of Figures: 3. No. of Tables: 1. No. of References: 22.

KEYWORDS generalized autoregression model; negative binomial; overdispersion; prediction; serial association

INTRODUCTION

The goal of the present paper is to apply methods existing in conditional inference to make predictions in time series of non-normal data. We shall focus on the particular case of overdispersed count data, although the presentation is sufficiently general to be applied in other contexts. The data set of interest concerns the growth of three closed colonies of *Paramecium aurelium* in a nutritive medium on a 20-day period. The observed counts are plotted in Figure 1. Details concerning the experiment can be found in Diggle (1990, p. 8). We have (artificially) truncated one of the series at day 10. We propose to build likelihood prediction envelopes for the discarded part of the series and to check that the actually observed values fall in these intervals.

* Correspondence to: Philippe Lambert, Faculty of Economics, Business and Social Sciences, University of Liège, Belgium

CCC 0277–6693/97/030195–13$17.50

Received March 1996
Accepted August 1996

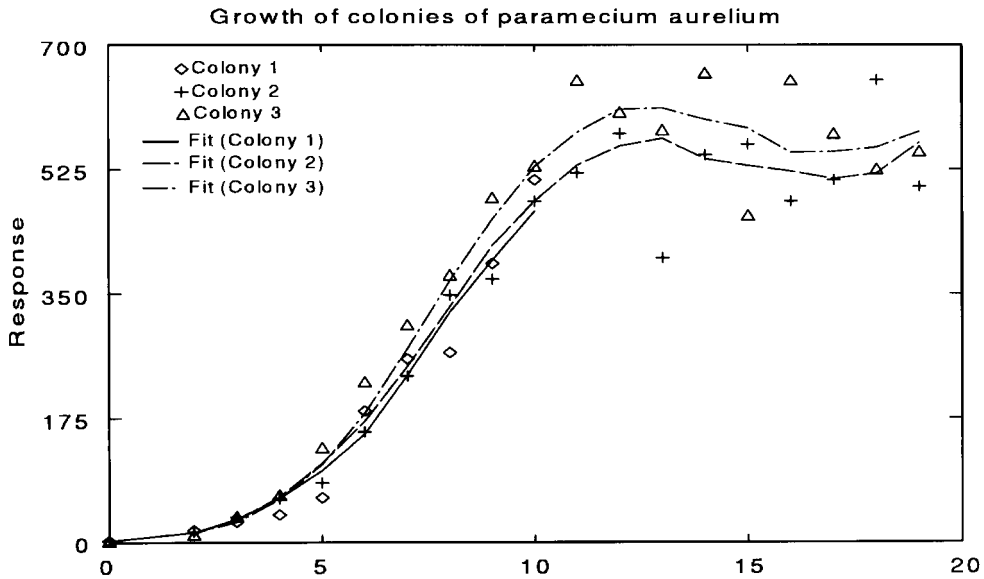

Figure 1. Growth of three closed colonies of *Paramecium aurelium* in a nutritive medium (Gause, 1934): fitted profile and data

The first step in the forecasting procedure is to build a likelihood for the observed data. Various techniques have been proposed in the literature to model series of counts. Most of them assume that the data are equally spaced and seem to suggest that extension to continuous time can be done easily, although this is often not true. The model considered for the observed part of the series is based on the generalized autoregression model (GARM) (Lambert, 1996a) originally used with series of non-negative data. This flexible tool, taking serial association into account, can be used to model virtually any type of nonnormal series of data. It is presented more specifically in a count data context in the next section.

A rewriting of the generalized autoregression model as a classical generalized linear model (GLM) is then proposed. This enables us to use the classical iterated weighted least squares (IWLS) algorithm to estimate the linear parameters in the model when the two serial association parameters (and possibly scale or shape parameters) are fixed. Only a non-linear optimizer is then required to obtain the maximum likelihood estimates (MLEs) of these parameters.

Our goal is to construct a likelihood function $L(z)$ where the datum z to be predicted plays the same role as a parameter of interest in an inference procedure. The most likely value $\hat{z}$ such that

$$\max_z L(z) = L(\hat{z})$$

would then be our point prediction. Similarly, $p\%$ likelihood prediction interval could be defined as

$$\{z : L(z)/L(\hat{z}) > p\%\}$$

The classical likelihood extension in this context is based on the *relative likelihood* described in Fisher (1959, pp. 128–33) and in Lejeune and Faulkenberry (1982). It consists of two parts: the first is the contribution to the likelihood made by the observed data, as provided, for example, by the GARM; the second is related to the observation to come. The profile likelihood

$$L(z) = L(z; \hat{\boldsymbol{\theta}} = (\hat{\rho}, \hat{\phi}, \hat{\boldsymbol{\beta}}, \hat{\upsilon})) = \max_{\boldsymbol{\theta}} L(z, \boldsymbol{\theta})$$

obtained by replacing the serial association (ρ and ϕ; see the next section), the regression ($\boldsymbol{\beta}$) and other nuisance (υ) parameters by their respective MLEs in the relative likelihood $L(z, \boldsymbol{\theta})$, can be used to make inferences about the unobserved datum z. One major criticism of this approach is that it does not take into account the uncertainty attached to each of the estimated parameters. The profile likelihood just assumes that the estimate $\hat{\boldsymbol{\theta}}$ of the nuisance parameters are the true (or 'population') values. A consequence of this in our setting is an underestimated width for likelihood prediction intervals. One way round this difficulty (Hinkley, 1979) is to condition on the maximum likelihood estimates $\hat{\boldsymbol{\theta}}$ which have a given distribution. The resulting conditional density can be approximated using the modified profile likelihood (see the third section; Lindsey, 1996, p. 112). There is a considerable literature on the subject which often relies on known analytic forms for the MLEs or on a substantial reduction of the data to sufficient statistics (Butler, 1986, 1989; Kalbfleisch and Sprott, 1970; Hinkley, 1979; Bjornstad, 1990; and, indirectly, Barndorff-Nielsen, 1983, 1993). For a review of prediction techniques based on the likelihood, see Bjornstad (1990). Unfortunately, with distributions outside the exponential family, problems related to the evaluation of jacobians in the modified profile likelihood arise. The approximate predictive likelihood $L_{\mathrm{B}}(z)$ proposed by Butler (see the third section 3; Butler, 1986, Rejoinder) was found more flexible in this context. Note also that the restriction to a linear form for the mean response in the generalized autoregression model (see the next section; Lambert, 1996a) simplifies the procedure. But even with such simplifications, a simple plot of the modified predictive likelihood is computationally intensive because it requires the numerical evaluation of the new maximum likelihood estimates $\hat{\boldsymbol{\theta}}$ for each value of z considered. From this plot of z, $p\%$ (Butler approximate) prediction intervals

$$\{z : L_{\mathrm{B}}(z)/L_{\mathrm{B}}(\hat{z}) > p\%\}$$

can be derived where $\hat{z}$ is the point prediction maximizing $L_{\mathrm{B}}(z)$. The correction to the conventional (Fisher) relative likelihood prediction intervals, not surprisingly, results into a small bias correction and slightly wider intervals. Note that this last effect is not that marked in our example. But the correction obtained by the technique might be more substantial with shorter series and smaller observed counts, or if all series had been truncated at day 10.

THE GENERALIZED AUTOREGRESSION MODEL

In this section we propose to review the ideas underlying the generalized autoregression model (GARM). It was originally considered in a completely different setting in Lambert (1996a). The problem of interest was then the modelling of series of dog blood parameters observed at

irregularly spaced time points. The response variables were typically non-negative and were shown to require distributions other than the usual normal and log-normal alternatives. This led us to develop a general methodology for modelling series of non-normal data. Here, we propose to illustrate the technique on the count data of the previous section. Instead of the traditional Poisson distribution, we have chosen the negative binomial which can deal with overdispersion. Basically a chosen transform of some location parameter (such as the mean response) is expressed as a function of the explanatory variables (here, time t_{ij}) plus one extra term accounting for the serial association. Mathematically speaking, denote by $f(y_{ij} \mid \mu_{ij}, \{y_{i1}, \ldots, y_{i,j-1}\}; \boldsymbol{\nu}_i)$ a distribution (not necessarily a member of the exponential family) for the response (conditional on past observations) where y_{ij} and μ_{ij}, respectively, denote the response and some location parameter for unit $i (i = 1, \ldots, I)$ at the jth $(j = 1, \ldots, n_i)$ sampling time t_{ij}. The vector $\boldsymbol{\nu}_i$ stands for other parameters such as scale or shape parameters. Note that $\boldsymbol{\nu}_i$ can include parameters common to all the series. Hence the vector of nuisance parameters might just be a scalar independent of i, as in the example below. More specifically, in our example, we have considered the following mean parameterization for the negative binomial

$$\Pr(Y_{ij} = y_{ij}) = \frac{\Gamma(y_{ij} + \nu)}{y_{ij}!\Gamma(\nu)} \left(\frac{\mu_{ij}}{\nu + \mu_{ij}}\right)^{y_{ij}} \left(\frac{\nu}{\nu + \mu_{ij}}\right)^{\nu} = f(y_{ij} \mid \{y_{i1}, \ldots, y_{i,j-1}\}, \mu_{ij}, \nu) \tag{1}$$

where

$$\begin{cases} i \in \{1, 2, 3\}, n_1 = n_2 = 19, n_3 = 10 \\ \nu > 0, y_{ij} \in \{0, 1, 2, \ldots\}, \mathrm{E}(Y_{ij}) = \mu_{ij} > 0 \end{cases}$$

Denote by

- $g(.)$ the desired transformation of the mean, or 'link function' (as in the GLM terminology). One could, for example, decide to model the log of the mean in the Poisson or the negative binomial distribution.
- r_{ij}, the residual for unit i at time t_{ij} on the g-scale, defined by

$$\begin{cases} r_{ij} = g(y_{ij}) - \mathbf{x}_{ij}^T \boldsymbol{\beta}_{ij} \\ r_{i0} = 0 \end{cases}$$

  where $\boldsymbol{\beta}$ stands for the regression parameter and $\mathbf{x}_{ij}$ for the vector of regressors at time t_{ij}. This definition for the residual is rather arbitrary, as pointed out by Lindsey (1993, p. 56). Note also that non-linear forms can be considered for the systematic part of the model. In our specific case, if we decide to model the log of the mean as a polynomial function of degree q (say), then one would set

$$g(\cdot) = \log(\cdot), \mathbf{x}_{ij}^T = (1, t_{ij}, \ldots, t_{ij}^q), \boldsymbol{\beta}_{ij} \in \mathbb{R}^{q+1}$$

  The residual r_{ij} is then the 'unexplained' part of the response on the log-scale. It is illustrated on Figure 2 in the case $q = 1$. Note that this plot is typical when dealing with serially associated observations: residuals observed at close time points tend to be of the same sign and order.

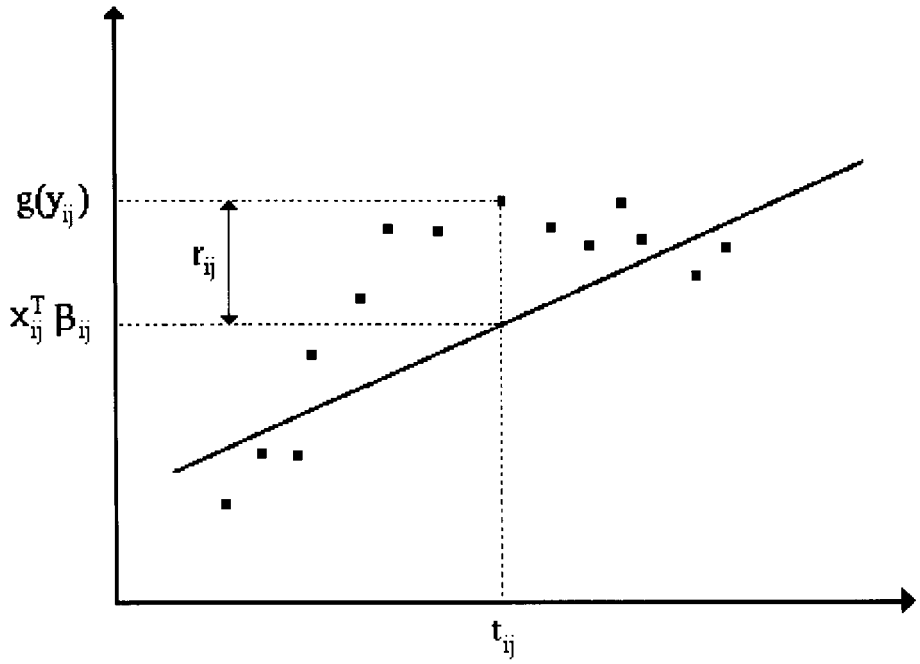

Figure 2. Typical plot of serially associated data against time

- r^c_{ij} the *cumulated residual* for unit i at time t_{ij} on the g-scale, defined by

$$\begin{cases} r^c_{ij} = e^{-\phi \Delta t_{ij}} r^c_{i,j-1} + r_{ij} \\ r^c_{ij} = 0 \\ w^c_{ij} = e^{-\phi \Delta t_{ij}} w^c_{i,j-1} + 1 \\ w^c_{i0} = 0 \end{cases}$$

  where $\Delta t_{ij} = t_{ij} - t_{i,j-1}$ and $0 < \phi$. Thus r^c_{ij} is a weighted sum of past residuals, with weights defined by w^c_{ij}. Then, the idea is to use $r^c_{i,j-1}$ as a forecast for the residual at the following observation time t_{ij}. Note that ϕ will be modelling the relative importance of former residuals in the prediction of this last quantity.

Then we model the g-transform of the location parameter as

$$g(\mu_{ij}) = \mathbf{x}^T_{ij}\boldsymbol{\beta} + e^{-\rho \Delta t_{ij}} \frac{r^c_{i,j-1}}{w^c_{i,j-1}} \tag{2}$$

where $0 < \rho$.

The first term in equation (2) corresponds to the usual modelling of the covariate influence on the response in generalized linear models. This is the solid line in Figure 2. The second term is proportional to the last cumulated residual that we hope to be close to the yet unobserved residual r_{ij}. Note that this residual correction is decreasing with the time elapsed (Δt_{ij}) since it

has been observed. A large value for ρ makes the second term meaningless, giving back the independence model. We refer the reader to Lambert (1996a) for more details.

We now propose to formulate the GARM in such a way that it can be handled by software like GLIM, S-Plus or SAS, if one restricts attention to distributions in the exponential family and to linear models for the systematic part. Consider the GARM model in equation (2) with the same notation as above. Then one can show that

$$\begin{cases} g(\mu_{i1}) = \mathbf{x}_{i1}^T\boldsymbol{\beta} \\ g(\mu_{i2}) = \mathbf{x}_{i2}^T\boldsymbol{\beta} + \mathrm{e}^{-\rho\Delta t_{i2}}[g(y_{i1}) - \mathbf{x}_{i1}^T\boldsymbol{\beta}] \\ \qquad = [\mathbf{x}_{i2} - \mathrm{e}^{-\rho\Delta t_{i2}}\mathbf{x}_{i1}]^T\boldsymbol{\beta} + \mathrm{e}^{-\rho\Delta t_{i2}}g(y_{i1}) \\ g(\mu_{i3}) = \mathbf{x}_{i3}^T\boldsymbol{\beta} + \mathrm{e}^{-\rho\Delta t_{i3}}\dfrac{[g(y_{i2} - \mathbf{x}_{i2}^T\boldsymbol{\beta}] + \mathrm{e}^{-\phi\Delta t_{i2}}[g(y_{i1}) - \mathbf{x}_{i1}^T\boldsymbol{\beta}]}{1 + \mathrm{e}^{-\phi\Delta t_{i2}}} \\ \qquad = \left[\mathbf{x}_{i3} - \mathrm{e}^{-\phi\Delta t_{i3}}\dfrac{\mathbf{x}_{i2} + \mathrm{e}^{-\phi\Delta t_{i2}}\mathbf{x}_{i1}}{1 + \mathrm{e}^{-\phi\Delta t_{i2}}}\right]^T\boldsymbol{\beta} + \mathrm{e}^{-\rho\Delta t_{i3}}\dfrac{g(y_{i2}) + \mathrm{e}^{-\phi\Delta t_{i2}}g(y_{i1})}{1 + \mathrm{e}^{-\phi\Delta t_{i2}}} \end{cases}$$

More generally, if we define as an *offset*, some fixed regression parameter in a regression model, we have

$$g(\mu_{ij}) = \mathbf{x}_{ij}^{\mathrm{d}T}\boldsymbol{\beta} + \mathrm{offset}_{ij} \tag{3}$$

where

$$\begin{cases} \mathbf{x}_{ij}^{\mathrm{d}} = \mathbf{x}_{ij} - \mathrm{e}^{-\rho\Delta t_{ij}}\dfrac{\mathbf{x}_{i,j-1}^{\mathrm{c}}}{w_{i,j-1}^{\mathrm{c}}} \\ \mathbf{x}_{ij}^{\mathrm{c}} = \mathbf{x}_{ij} + \mathrm{e}^{-\rho\Delta t_{ij}}\mathbf{x}_{i,j-1}^{\mathrm{c}} \\ \mathbf{x}_{ij}^{\mathrm{c}} = \mathbf{x}_{ij} \\ \gamma_{ij}^{\mathrm{c}} = g(y_{ij}) + \mathrm{e}^{-\rho\Delta t_{ij}}\gamma_{i,j-1} \\ \gamma_{i1}^{\mathrm{c}} = g(y_{i1}) \\ \mathrm{offset}_{ij} = \mathrm{e}^{-\rho\Delta t_{ij}}\dfrac{\gamma_{i,j-1}^{\mathrm{c}}}{w_{i,j-1}^{\mathrm{c}}} \end{cases} \tag{4}$$

From equation (3), we conclude that, for fixed values of ρ and ϕ, the GARM can be expressed as a GLM where, for each unit, and some observation time t_{ij},

- Any row $\mathbf{x}_{ij}^{\mathrm{d}T}$ of the design matrix can be expressed as the difference between the original design matrix row $\mathbf{x}_{ij}^T$ and some weighted average of design matrix rows from previous observation times.
- An offset, which is a weighted average of the last observation on the g-scale and the previous g-observations on that unit, is introduced.

This formulation will be particularly useful below to compute modified or approximate predictive likelihoods. We give more details in the Appendix on how to use GLM software to fit a GARM to the illustrative data set.

LIKELIHOOD PREDICTION ENVELOPES

Let us assume that a suitable model has been found for the observed data and that the contribution

$$f(y_{ij} \,|\{y_{i1}, \ldots, y_{i,j-1}\}; \mu_{ij}, \mathbf{v}_i)$$

to the likelihood for each observation is available.

We shall restrict our attention to likelihood based prediction methods. Suppose that we want to predict an extra observation z on the kth (say) series at some given time $t = t_{k,n_k+1}$. As already mentioned in the introduction, one naive but simple way to make predictions is to consider the profile predictive likelihood (Fisher, 1959, pp. 128–33, Lejeune and Faulkenberry, 1982):

$$L(z \,|\, \mathbf{y}_1, \ldots, \mathbf{y}_I) = f(z \,|\, \mathbf{y}_k; \hat{\mu}^{(z)}_{k,n_k+1}, \hat{\mathbf{v}}^{(z)}_k) \prod_i \prod_j f(y_{ij} \,|\{y_{i1}, \ldots, y_{i,j-1}\}; \hat{\mu}^{(z)}_{ij}, \hat{\mathbf{v}}^{(z)}_i)$$

where z stands for the observation to come on unit k (which now plays the role of the parameter of interest in the likelihood), $\mathbf{y}_i$ denotes the set of observations on unit i, and $\hat{\mu}^{(z)}_{ij}$ and $\hat{\mathbf{v}}^{(z)}_i$, respectively, denote the MLEs of μ_{ij} and $\mathbf{v}_i$ given $\{y_{i1}, \ldots, y_{in}, z\}$ and the observations on the other series. The symbol (z) as superscript is used to make a distinction between the MLEs computed from the likelihood including the contribution of the unobserved z datum and the 'classical' MLEs computed using only the observed data contribution to the likelihood. More precision is achieved by estimating the model parameters using a likelihood based on all the series instead of using the sole likelihood contribution from the series of interest. Note that the first element in the profile predictive likelihood is related to the observation to come, whereas the others are the contribution of the observed data.

As already mentioned, assuming the unknown parameters to be fixed at their MLEs is not realistic (Butler, 1986): hence the need for likelihood methods based on distributions conditional to the maximum likelihood estimates of the nuisance parameters. Deriving such quantities is not an easy task and except in very special circumstances one has to approximate the conditional distribution. The so-called Barndorff-Nielsen (1983) p^*-formula provides an approximation to the distribution of MLEs given an ancillary statistic. The required (approximate) conditional distribution can then be derived yielding

$$L^*(z \,|\, \mathbf{y}_1, \ldots, \mathbf{y}_I) = L(z \,|\, \mathbf{y}_1, \ldots, \mathbf{y}_I) |\, J^{(z)}(\hat{\boldsymbol{\theta}}^{(z)})|^{-1/2} \left| \frac{\partial\, \hat{\boldsymbol{\theta}}}{\partial\, \hat{\boldsymbol{\theta}}^{(z)}} \right| \tag{5}$$

where $\boldsymbol{\theta}^T = (\boldsymbol{\beta}^T, \mathbf{v}_1^T, \ldots, \mathbf{v}_1^T,)$ stands for all the parameters in the model such as the regression parameters $\boldsymbol{\beta}$ defining μ_{ij} and the nuisance parameters; $J^{(z)}(\hat{\theta}^{(z)})$ is the observed information matrix about $\boldsymbol{\theta}$ computed at the MLEs for a given value of the unobserved data z, i.e.

$$J^{(z)}(\hat{\boldsymbol{\theta}}^{(z)}) = \frac{\partial^2 \log f(z \,|\, \mathbf{y}_k; \boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}\, \partial\, \boldsymbol{\theta}^T} \bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(z)}} + \sum_i \sum_j \frac{\partial^2 \log f(y_{ij} \,|\{y_{i1}, \ldots, y_{i,j-1}\}; \boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}\, \partial\, \boldsymbol{\theta}^T} \bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(z)}}$$

Unfortunately the last factor in equation (5) is particularly tedious to compute because it requires the analytical expression of the MLEs, which, outside the exponential family, are usually

impossible to determine. In such settings, Butler (1986, Rejoinder) proposes an approximate predictive likelihood that can be evaluated with any type of distribution because it only requires the maximum likelihood estimates of the nuisance parameters based on $\{\mathbf{y}_1^T, \ldots, \mathbf{y}_I^T, z\}$:

$$L_B(z \mid \mathbf{y}_1, \ldots, \mathbf{y}_I) = L(z \mid \mathbf{y}_1, \ldots, \mathbf{y}_I) \mid J^{(z)}(\hat{\boldsymbol{\theta}}^{(z)})\mid^{1/2} \left| H(\hat{\boldsymbol{\theta}}^{(z)}) H^T(\hat{\boldsymbol{\theta}}^{(z)}) \right| \tag{6}$$

with

$$H^{(z)}(\hat{\boldsymbol{\theta}}^{(z)}) = \frac{\partial^2 \log f(z \mid \mathbf{y}_k; \boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}\, \partial\{\mathbf{y}_1^T, \ldots, \mathbf{y}_I^T\}} \Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(z)}} + \sum_i \sum_j \frac{\partial^2 \log f(y_{ij} \mid \{y_{i1}, \ldots, y_{i,j-1}\}; \boldsymbol{\theta})}{\partial\, \boldsymbol{\theta}\, \partial\{\mathbf{y}_1^T, \ldots, \mathbf{y}_I^T\}} \Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(z)}}$$

As briefly explained in the original paper, this last formula for the conditional likelihood can be derived using a Taylor series expansion for the joint density of the observed and unobserved data about the nuisance parameter MLEs, and by dividing the whole by a normal approximation to the joint distribution of the nuisance parameters. Using the inverse of the Fisher information evaluated at $\hat{\theta}$ as an approximation to the covariance matrix of the above normal distribution simplifies the final formula, yielding equation (6).

As already mentioned in the introduction, the aim of this paper is to construct likelihood prediction envelopes. By a p% likelihood prediction envelope, we mean a succession of p% (Butler approximate) predictive likelihood intervals computed at the time points of interest. These likelihood intervals are obtained in the same way as with a traditional parameter likelihood (Kalbfleisch, 1985), the role of the parameter here being played by the unobserved quantity. We have not tried to compute *simultaneous* prediction intervals, to make an analogy with the frequentist simultaneous confidence intervals. In our view one is more interested by what is 'likely' to be observed in the future at one given time point independently of what the other predictions are. Plotting an envelope is more a way to summarize graphically a series of independent results than giving artificially related statements. However, a *simultaneous* approach is feasible, but this is technically far more difficult, particularly in a non-normal setting, because it requires the computation of N_p- (and thus possibly large) dimension normed likelihood regions (if one wants to predict N_p unobserved data). Finally, note that a method for modelling series of data observed at unequally spaced times is required because the time at which the prediction is made is (in practice) totally arbitrary. A second (but not rigorous) choice would be to proceed step by step by using a conditioning argument. The observation at time t_{k,n_k+1} could be predicted conditional on the last observation; the one at time t_{k,n_k+2} could be derived by conditioning on the first-step prediction; and so on (with accumulation of errors) until the time of interest has been reached. This can be useful when the point forecast is really what interests us, but this becomes far more complicated when prediction intervals are required.

APPLICATION

Applying the theory of the previous sections to construct a likelihood prediction envelope for the truncated part of the introduction data set is the subject of this section. The full data set has been studied by Diggle (1990, pp. 155) who proposed a quartic polynomial in time to model the observed growth curve. Lambert (1996b) uses a generalized form of the logistic growth curve

(Nelder, 1961, 1962) to take into account the asymptotic behaviour of the colony sizes and compares it with the quartic polynomial fit. It was noticed that both solutions fit equally well empirically in the observed time range, but that the generalized logistic form should be preferred because it is more sensible than polynomials to model biological mechanisms of growth.

However, because forecasts are made in the observed time range, and in order to illustrate the expression of the GARM as a GLM, we have decided to present the construction of the likelihood prediction envelopes with the quartic polynomial (and thus linear) model. Note that the same approach can be used with the generalized logistic form, but this would require the use of specially written FORTRAN of GAUSS (in our case) code to compute MLEs.

As already pointed out in Lambert (1996b), a negative binomial distribution seems to be more adapted than the Poisson alternative.

The influence of time as well as of any other explanatory variable on the mean response μ_{ij} on unit i at time t_{ij} can be modelled using the regression

$$\log(\mu_{ij}) = \mathbf{x}_{ij}^{\mathrm{d}^T}\boldsymbol{\beta} + \mathrm{offset}_{ij}$$

jointly with a GARM to take serial dependence into account. For a quartic polynomial in time, take $\mathbf{x}_{ij}^T = (1, t_{ij}, \ldots, t_{ij}^4)$; $\mathbf{x}_{ij}^{\mathrm{d}}$ can be deduced from equation (4). Note that the GARM appears in the regression equation through the offset and the transformed design matrix X^{d}. We refer the reader to the Appendix for details on the estimation procedure.

If the negative binomial parameter v and the serial association parameters ρ and ϕ were known, then one could simply compute the regression parameter MLEs using the IWLS algorithm. Taking a grid of values for the three unknown parameters might be one solution to determine the MLEs. In this example we have used the non-linear optimizing procedure PROC OPTMUM in GAUSS. The corresponding MLEs and fit are respectively displayed in Table I and Figure 1. The fitted profiles displayed on this figure are not smooth curves because corrections due to serial association were added to the polynomial contribution.

Table I. Parameter MLEs computed on the *Paramecium Aurelium* data set using a logistic regression jointly with a GARM

Par.	Est.
υ	63.38
Serial association Par.	
ϕ	0.2786
ρ	0.7685
Regression Par.	
β_0	0.6804
β_1	1.104
β_2	-6.493×10^{-2}
β_3	7.550×10^{-4}
β_4	2.260×10^{-5}

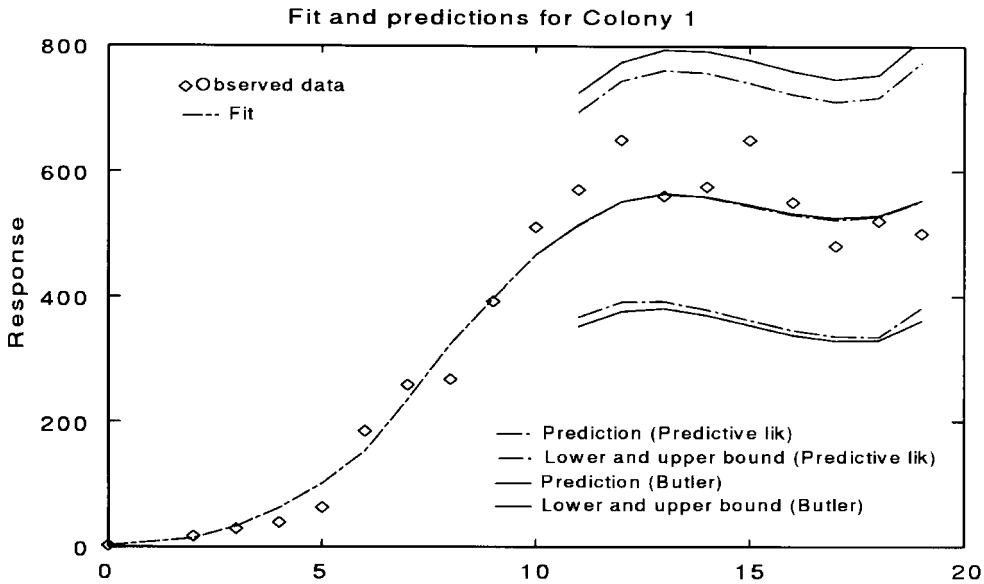

Figure 3. Growth of three closed colonies of *Paramecium aurelium* in a nutritive medium (Gause, 1934): predictions and likelihood prediction envelope for the artificially truncated series

The next step is the computation of the Butler approximate predictive likelihood for various values of z at the time point of interest. This can be done in three steps:

(1) Given the likelihood based on the observed data and z, compute the MLEs of the eight parameters in the model (see Table I).
(2) Compute the Jacobian $J^{(z)}(\hat{\boldsymbol{\theta}}^{(z)})$ and the matrix $H(\hat{\boldsymbol{\theta}}^{(z)})$ at the MLEs from step 1.
(3) Compute the approximate predictive likelihood using equation (6) together with the results of the first two steps.

In our example, the procedure has been simplified by only conditioning on the regression parameter estimates. This reduces the dimension of the Jacobian and of the H matrices from eight to five. Of course, it is not necessary to use numerical methods to compute the J and H matrices: analytic forms are easy to derive. This point is essential to ensure a reasonable rapidity to the procedure. Indeed one has to repeat the three above-described steps for different values of z to first determine the forecast $\hat{z}$ (which maximizes the approximate predictive likelihood) at the time point of interest. Once $\hat{z}$ is known, the predictive likelihood is rescaled by dividing it by its maximum value. The 10% (say) predictive likelihood interval can then be determined. This can be done using, for example, the secant method for determining the zeros of a function. This method has the advantage of not requiring the derivative of the predictive likelihood (at the cost of a function evaluation).

The corresponding results for the *Paramecium aurelium* data set are displayed in Figure 3. Two approaches were used to compute the 10% likelihood prediction envelope. The first was based on the Fisher predictive likelihood which assumes that the parameters are known (and equal to their MLEs). The second method is the one described above. As can be seen from Figure 3, there is a small bias correction and the likelihood prediction envelope derived using the Butler

approximation is slightly wider, reflecting the extra uncertainty on the regression parameters values. It would probably be even wider if all eight parameters were used when conditioning. The correction with respect to the Fisher method is not very important in this setting. We would expect larger corrections with smaller counts and shorter series. Finally, note that the actually observed data fall well within the 10% likelihood prediction envelope (whatever the chosen method).

DISCUSSION

In this paper we have shown how to express the generalized autoregression model (Lambert, 1996a) or GARM as a GLM when a linear model for a given transform of the location parameter is chosen. This enables the use of the IWLS algorithm which is available in most statistical software. However, a non-linear optimizer has to be used to determine the MLEs of the serial association, shape and scale parameters. The example considered in the previous section involves only three such parameters.

The Butler approximate predictive likelihood (Butler, 1986, Rejoinder) has been used to compute likelihood prediction envelopes. The resulting likelihood prediction intervals were shown to be slightly wider than the Fisher (1959, pp. 128–33) intervals, although the size of the correction might be more important in other settings. Moreover, a small bias correction was pointed out. Other types of correction to the usual profile likelihood are available in the literature (Barndorff-Nielsen, 1983; Cox and Reid, 1987; Davison, 1986, 1987; Fraser and Reid, 1989), but they are often either uncomputable outside the exponential family or unadapted to a prediction problem. For example, the Cox and Reid (1987) adjusted profile likelihood requires the prediction (seen as a parameter in the likelihood) to be information orthogonal to the regression parameters, which, in a prediction context, is not very sensible. Thus the choice of the Butler approximate predictive likelihood was more pragmatic than theoretical.

APPENDIX

Here we show how to use GLM software to fit a negative binomial autoregression model to a data set. One common way to fit the negative distribution is to take advantage of the similarity between its likelihood and the binomial likelihood. Indeed, rewrite equation (1)

$$\Pr(Y_{ij} = y_{ij}) = \frac{\Gamma(y_{ij} + v)}{y_{ij}!\Gamma(v)} \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{v}$$

and compare this formula with the binomial likelihood contribution

$$\Pr(Y_{ij} = y_{ij}) = \frac{n_{ij}!}{y_{ij}!(n_{ij} - y_{ij})!} \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{n_{ij} - y_{ij}}$$

for an observed proportion y_{ij}/n_{ij} where n_{ij} is some number to be determined. Setting

$$n_{ij} = y_{ij} + v$$

and fitting (for a fixed value of v) a binomial distribution to the proportions

$$\{y_{ij}/n_{ij} : 1 < i < I, 0 < j < n_i\}$$

we get the same conditional (on v) MLEs $\hat{\pi}_{ij}$ as if we had maximized the negative binomial likelihood directly over π_{ij}. Of course, it is possible to include explanatory variables as in the autoregression model of equation (3). This can be reformulated as a binomial ('logistic') regression using the relation

$$\mathrm{E}(Y_{ij}) = \mu_{ij} = v\frac{\pi_{ij}}{1-\pi_{ij}}$$

Equation (3) then becomes

$$\log\frac{\pi_{ij}}{1-\pi_{ij}} = \mathbf{x}_{ij}^{\mathrm{d}^T}\boldsymbol{\beta} + \mathrm{offset}_{ij}^{'}$$

where

$$\mathrm{offset}_{ij}^{'} = \mathrm{offset}_{ij} - \log(v)$$

Using this trick and a GLM package, the regression parameter $\boldsymbol{\beta}$ can be estimated for fixed values of v and of the serial association parameters ρ and ϕ (present in $\mathbf{x}_{ij}^{\mathrm{d}}$ and offset_{ij}).

REFERENCES

Barndorff-Nielsen, O. E., 'On a formula for the distribution of the maximum likelihood estimator', *Biometrika*, **70** (1983), 343–65.

Barndorff-Nielsen, O. E., 'Adjusted likelihood inference about interest parameters', *Theory Probab. Appl.,* **38** (1993), 179–93.

Bjornstad, J. F., 'Predictive likelihood: a review', *Statistical Science*, **5** (1990), 242–65.

Butler, R. W., 'Predictive likelihood inference with applications (with discussion)', *Journal of the Royal Statistical Society*, **B48** (1986), 1–38.

Butler, R. W., 'Approximate predictive pivots and densities', *Biometrika*, **76** (1989), 489–501.

Cox, D. R. and Reid, N., 'Parameter orthogonality and approximate conditional inference', *Journal of the Royal Statistical Society*, **B49** (1987), 1–39.

Davison, A. C., 'Approximate predictive likelihood', *Biometrika*, **73** (1986), 323–32.

Davison, A. C., 'Discussion on Parameter orthogonality and approximate conditional inference (by D. R. Cox and N. Reid)', *Journal of the Royal Statistical Society*, **B49** (1987), 28–9.

Diggle, P. J., *Times Series. A biostatistical introduction*, Oxford: Oxford University Press, 1990.

Fisher, R. A., *Statistical Methods and Scientific Inference*, Edinburgh: Oliver and Boyd, 1959.

Fraser, D. A. S. and Reid, N., 'Adjustments to profile likelihood', *Biometrika*, **76** (1989), 477–88.

Gause, G. F., *The Struggle for Existence*, Baltimore, MD: Williams and Williams, 1934.

Hinkley, D., 'Predictive likelihood', *The Annals of Statistics*, **79** (1979), 718–28.

Kalbfleisch, J. D., *Probability and Statistical Inference*. Volume 2: *Inference*; New York: Springer-Verlag, 1985.

Kalbfleisch, J. D. and Sprott, D. A., 'Application of likelihood methods to models involving large numbers of parameters', *Journal of the Royal Statistical Society*, **B32** (1970), 175–208.

Lambert, P., 'Modelling irregularly sampled profiles of nonnegative dog triglyceride responses under different distributional assumptions', *Statistics in Medicine*, **15** (1996a), 1695–1708.

Lambert, P., 'Modelling of non-linear growth curve on series of correlated count data measured at unequally spaced times: a full likelihood based approach', *Biometrics*, **52** (1996b), 50–55.

Lejeune, M. and Faulkenberry, G. D., 'A simple predictive density function', *Journal of the American Statistical Association*, **77** (1982), 654–9.

Lindsey, J. K., *Models for Repeated Measurements*, Oxford: Oxford Statistical Science Series, 1993.

Lindsey, J. K., *Parametric Statistical Inference*, Oxford: Oxford Statistical Science Series, 1996.

Nelder, J. A., 'The fitting of a generalization of the logistic curve', *Biometrika*, **17** (1961), 89–100.
Nelder, J. A., 'An alternative form of a generalized logistic equation', *Biometrics*, **18** (1962), 614–16.

Author's biography:
Philippe Lambert is researcher in the Faculty of Economics, Business and Social Sciences at the University of Liège, Belgium. He received his PhD from the Limburgs Universitair Centrum, Diepenbeek (Belgium) in 1995. His current research interests include the modelling of non-normal longitudinal data, growth curve, stable processes and discrete data.

Author's address:
Philippe Lambert, Faculty of Economics, Business and Social Sciences, University of Liège, Bd du Rectorat, 7 (B31), B-4000 Liège, Belgium.

