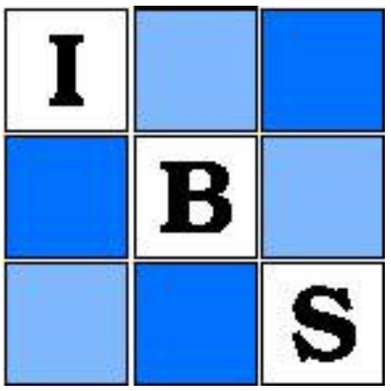

BIOMETRICS 52, 50–55
March 1996

Modeling of Nonlinear Growth Curve on Series of Correlated Count Data Measured at Unequally Spaced Times: A Full Likelihood Based Approach

Philippe Lambert

Faculté EGSS, University of Liège,
Bd du Rectorat, 7 (B31), B-4000 Liège, Belgium

SUMMARY

A "robust" version of the gamma-Poisson model (Lambert, P., 1996, *Applied Statistics,* in press) for series of count data observed at unequally spaced times is used to analyze the growth of three closed colonies of *Paramecium aurelium* in a nutritive medium (Diggle, P. J., 1990, *Time Series. A Biostatistical Introduction*) where successive sample counts within each replicate are likely to be statistically dependent. A generalized form of the logistic growth curve (Nelder, J. A., 1961, *Biometrika* **17,** 89–100; 1962, *Biometrics* **18,** 614–616) further developed by Heitjan (1991, *Statistics in Medicine* **10,** 1075–1088; 1991, *Journal of the American Statistical Association* **86,** 891–898) and including the Mitscherlich, Gompertz, logistic, and exponential forms as well-known members, was chosen to model the response profile. Comparisons with other (possibly nonnested) models are made using the Akaike criterion (Akaike, H., 1973, in *Second International Symposium on Inference Theory Petrov*).

1. Introduction

In this paper we present a model that should be useful to anyone dealing with nonlinear growth curve data recorded as series of counts measured at unequally spaced time points.

The data set of interest (in Table 1) concerns the growth of three closed colonies of *Paramecium aurelium* in a nutritive medium. The observed counts are plotted in Figure 1. Details concerning the experiment can be found in Diggle (1990, p. 8). Our goal is to build a model for the profile of the mean number of individuals.

The main difficulty is the specification of the autocorrelation structure. Any two measurements performed on the same observational unit should be closely related if they are made at close points in time. A minimal requirement is that the autocorrelation should be a decreasing function of time.

Many authors have proposed a solution to similar problems under the normal assumption. An elegant method of analysis based on the Kalman filter is given by Jones and Ackerson (1990), Jones and Boadi-Boateng (1991), and Jones (1993). These techniques usually assume a multivariate normal distribution for the data measured on the same unit, the measurements from two different series being assumed independent. The covariance matrix is given nonzero off-diagonal elements decreasing to zero with the time lag between the two corresponding observation times.

Unfortunately, those approaches cannot easily be extended to deal with nonnormal series of data in a likelihood-based approach. There are two main reasons for this. First, there are no simple methods to handle multivariate distributions for nonnormal data. Second, even if we had one, specification of the first- and second-order moments would not completely define the dependence structure.

The technique that we propose is an extension of the gamma-Poisson model (Harvey and Fernandes, 1989; Lindsey and Lambert, 1995; Ord, Fernandes, and Harvey, 1993) adapted to deal with

Key words: Autoregression; Discount parameter; Dynamic generalized linear model; Gamma-Poisson model; Growth curve; Kalman filter; Heterogeneity; Longitudinal data; Overdispersion; Repeated measurements.

unequally spaced observations by Lambert (1996). We briefly present the underlying argument again to enable a clear understanding of Section 2.

Table 1
Growth of three closed colonies of Paramecium aurelium in a nutritive medium (Diggle, 1990, p. 8)

Day	0	2	3	4	5	6	7	8	9	10
Sample 1	2	17	29	39	63	185	258	267	392	510
Day	11	12	13	14	15	16	17	18	19	
Sample 1	570	650	560	575	650	550	480	520	500	
Day	0	2	3	4	5	6	7	8	9	10
Sample 2	2	15	36	62	84	156	234	348	370	480
Day	11	12	13	14	15	16	17	18	19	
Sample 2	520	575	400	545	560	480	510	650	500	
Day	0	2	3	4	5	6	7	8	9	10
Sample 3	2	11	37	67	134	226	306	376	485	530
Day	11	12	13	14	15	16	17	18	19	
Sample 3	650	605	580	660	460	650	575	525	550	

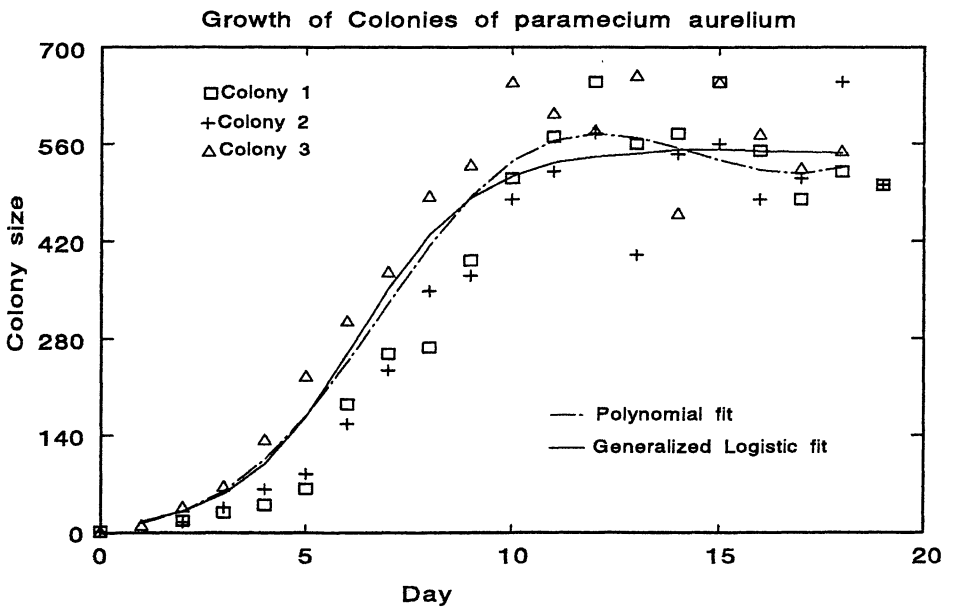

Figure 1. Growth of three closed colonies of *Paramecium aurelium* in a nutritive medium (Diggle, 1990, p. 8): fitted profiles and data.

Suppose that we observe several independent series of counts $\{y_{i1}, \ldots, y_{iJ_i}\}$ on I units ($i = 1, \ldots, I$) at unequally spaced times $\{t_{i1}, \ldots, t_{iJ_i}\}$ together with a set of covariates $\{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i}\}$. One possible model for such data is the gamma-Poisson model. The value y_{ij} is seen as generated by a Poisson process with a time-dependent mean $\exp(\beta_{ij}^T \mathbf{x}_{ij})$. One can release the linearity assumption for the part of the mean involving the covariates. We denote the resulting mean by $\lambda_{ij}\mu(\mathbf{x}_{ij})$. The nonrandom and possibly nonlinear part $\mu(\mathbf{x}_{ij})$ models the influence of covariates. Suitable forms for it in a growth curve context are described in Section 3. It can be defined to model the profile, the influence of continuous explanatory variables such as the dose of a given medicine, or the influence of an indicator variable. Note that $\log(\lambda_{ij})$ is the residual part of the mean at time t_{ij} for unit i on the log scale that is not modeled by regressors. If the data are autocorrelated, the values of two residuals observed at close time points on the same unit should be closely related. One way to allow for this feature is to give, at any time point, a distribution to λ_{ij} and to ensure that

it is evolving in a "smooth" way with time. Fahrmeir (1992) used a normal prior, but, for analytic simplicity, we choose a conjugate gamma distribution. Once an observation is available at time t, the corresponding prior distribution (which is the posterior at the previous observation time) has to be updated (using Bayes' theorem) to account for the observed residual that it has generated. The further back in time the previous observation, the less weight the prior is given in building the corresponding posterior distribution. The parameter used for modeling the prior weight was shown to behave like a serial association parameter.

Finally, we propose different parametric profile expressions to describe nonlinear growth curves. Comparison of the generalized logistic growth curve (Nelder, 1961, 1962) with polynomial models is made by computing the AIC criterion (Akaike, 1973), which is most appropriate in a model selection context (Lindsey, 1994). The improvement brought by the (extended) gamma-Poisson (stochastic) model over the negative binomial (which assumes independence) is also explored.

2. Extending the Gamma-Poisson Model

Denote by $\mathcal{F}_{ij}$ ($\mathcal{F}_{i,j|j-1}$) the *filtration* or history of the responses for unit i up to (but not including) time t_{ij}. We will use the shorthand notation μ_{ij} for the nonrandom part $\mu(\mathbf{x}_{ij})$ of the mean.

Similar to Lambert (1996), we consider the gamma-Poisson model

$$\begin{aligned}(Y_{ij} \mid \lambda_{ij}, \mathbf{x}_{ij}, \mathcal{F}_{i,j|j-1}) &\sim \text{Poi}(\lambda_{ij}\mu_{ij}),\\ (\lambda_{ij} \mid \mathcal{F}_{i,j|j-1}) &\sim \text{G}(\kappa_{i,j|j-1}, \upsilon_{i,j|j-1}),\end{aligned}$$

where $p(\lambda_{ij} \mid \mathcal{F}_{i,j|j-1})$ denotes the gamma prior distribution (with Fisher information υ/κ at the mode κ) for the residual. That prior was computed from the last posterior using

$$\begin{aligned}\kappa_{i,j|j-1} &= \kappa_{i,j-1},\\ \upsilon_{i,j|j-1} &= \text{e}^{-\phi\Delta t_{ij}}\upsilon_{i,j-1},\end{aligned}$$

thereby expressing the loss of information during the time interval $\Delta t_{ij} = t_{ij} - t_{i,j-1}$. Note that the preceding specification of the prior enables us to cope with continuous observation times.

That prior is then updated using Bayes' theorem at any observation time t_{ij} to give account that it has generated y_{ij}. That yields a gamma posterior

$$(\lambda_{ij}|\mathcal{F}_{ij}) \sim \text{G}(\kappa_{ij}, \upsilon_{ij})$$

with

$$\kappa_{ij} = \kappa_{i,j|j-1} + \frac{y_{ij} - \kappa_{i,j|j-1}\mu_{ij}}{\upsilon_{i,j|j-1} + \mu_{ij}}, \tag{1}$$

$$\upsilon_{ij} = \upsilon_{i,j|j-1} + \mu_{ij}. \tag{2}$$

The resulting likelihood is a product over units of negative-binomial distributions from which maximum likelihood estimators (MLEs) for ϕ and μ_{ij} can be numerically computed (using Proc OPTMUM in GAUSS, for example).

Note that modeling of heterogeneity is at least partially accounted for by assuming a possibly different evolution of the residuals for two different units. One could also further condition on (the logarithm of) the first observation by using μ_{ij}.

We refer the reader to the afore mentioned paper for both technical and intuitive details.

We now propose to extend the preceding gamma-Poisson model by introducing two extra parameters:

1. It would be desirable to allow the gamma-Poisson model to include the simple Poisson distribution as a special case. That can be achieved by considering λ'_{ij} instead of λ_{ij}, where

$$p(\lambda'_{ij}|\mathcal{F}_{i,j|j-1}) \propto p(\lambda_{ij}|\mathcal{F}_{i,j|j-1})^{\delta}.$$

The gamma mixing distribution will reduce to a point if δ tends to zero and to a vague distribution if δ tends to infinity. The recursive procedure will be similar to preceding, but with equations (1) and (2) replaced by

$$\begin{aligned}\kappa'_{ij} &= \kappa'_{i,j|j-1} + \frac{y_{ij} - \kappa'_{i,j|j-1}\mu_{ij}}{\upsilon'_{i,j|j-1}\delta + \mu_{ij}},\\ \upsilon'_{ij} &= \upsilon'_{i,j|j-1} + \frac{\mu_{ij}}{\delta}.\end{aligned}$$

2. A further interesting parameter is $0 \leq \alpha \leq 1$, where

$$\begin{aligned} \kappa'_{ij} &= \kappa'_{i,j|j-1} + \alpha \frac{y_{ij} - \kappa'_{i,j|j-1}\mu_{ij}}{\upsilon'_{i,j|j-1}\delta + \mu_{ij}}, \\ \upsilon'_{ij} &= \upsilon'_{i,j|j-1} + \alpha \frac{\mu_{ij}}{\delta}. \end{aligned} \tag{3}$$

to reduce the sensitivity of the model to extreme observations. Values of α close to zero avoid to have a too quick correction (as implied by equation (3)) of the model toward an outlier, that is, an observation for which $y_{ij} - \kappa'_{i,j|j-1}\mu_{ij}$ is large.

Note that both parameters have substantially decreased the AIC of the model when applied to most of the data sets that we have examined so far. More important, both parameters were required to get a better fit than the negative binomial model (which assumes independence), even when applied to equally spaced count data. In other words, the Harvey and Fernandes (1989) model was not able to challenge the negative binomial with most of the data sets that we have manipulated in consulting, but we should remember that these two authors were working in a forecasting context.

3. Modeling Growth Curve

One striking feature of our data set is the stabilization of the colonies after about 10 days. Therefore, the systematic part of the model should tend to an asymptote as time passes. The model considered by Diggle (1990, p. 155) does not allow for this. He just considered a quartic polynomial in time for the mean number of individuals in each colony. The fit may be reasonable within the observed time span, but is not realistic for larger time values.

Nelder (1961, 1962) considered a generalization of the logistic growth curve, further developed by Heitjan (1991a, 1991b) and including the Mitscherlich, Gompertz, logistic, and exponential forms as well-known members. Heitjan (1991b) used that family of models to assess the effect of three multiple sclerosis treatments on the evolution of ACFR, a measure of autoimmunity. Inclusion of explanatory variables in the systematic part is illustrated.

The equation of those profiles is given by

$$\mu_{ij} = \mathrm{e}^{\kappa_2}\left[1 + (\mathrm{e}^{(\kappa_2-\kappa_1)\kappa_4} - 1)\mathrm{e}^{-\kappa_3 t_{ij}\mathrm{e}^{\kappa_2\kappa_4}}\right]^{-\frac{1}{\kappa_4}}, \qquad \kappa_4 \neq 0, \tag{4}$$

$$= \mathrm{e}^{\kappa_2 + (\kappa_1-\kappa_2)\mathrm{e}^{-\kappa_3 t_{ij}}}, \qquad \kappa_4 = 0. \tag{5}$$

It is the solution of the differential equation

$$\frac{d\mu_{ij}}{dt_{ij}} = \kappa_3 \mu_{ij}\left[d(\mathrm{e}^{\kappa_2}, \kappa_4) - d(\mu_{ij}, \kappa_4)\right],$$

where

$$\begin{aligned} d(\mu_{ij}, \kappa_4) &= \frac{\mu_{ij}^{\kappa_4} - 1}{\kappa_4}, \qquad \kappa_4 \neq 0, \\ &= \log(\mu_{ij}), \qquad \kappa_4 = 0. \end{aligned}$$

Note that $\kappa_1 = \log(\mu_{i0})$ is the initial condition and $\kappa_2 = \lim_{t_{ij}\to\infty} \log(\mu_{ij})$ the asymptote. The parameters κ_3 and κ_4 model the rate of growth. The parameter κ_4 also determines the type of the curve, varying from the Mitscherlich ($\kappa_4 = -1$) through the Gompertz ($\kappa_4 = 0$), and the logistic ($\kappa_4 = 1$) to the exponential ($\kappa_4 \to \infty$ and $d(\mathrm{e}^{\kappa_2}, \kappa_4) \to$ constant) (Heitjan, 1991b).

4. Results

Two families of models were considered. The first one, ignoring the longitudinal aspect of the data set, and thus ignoring autocorrelation between data on the same unit, assumes that the counts are distributed as a negative binomial. The second family is the gamma-Poisson model where autocorrelation is modeled using an empirical Bayes approach for the residuals on the log scale, that choice being suggested by the canonical link for Poisson data.

For each family of models, two types of systematic parts for the mean were fitted to the data:

- A degree four polynomial in time, as suggested by Diggle (1990, p. 155).
- The generalized logistic growth curve given by equations (4) and (5) (Nelder, 1961, 1962).

The parameter estimation of the gamma-Poisson model was performed into three steps. First, the MLEs for the regression parameters in the independence negative binomial model were computed. Then the serial association parameters ϕ, δ, and α of the gamma-Poisson model were estimated, the regression parameters being held fixed at their first step values. Finally the gamma-Poisson likelihood was maximized over both the serial association and regression parameters, starting values being given by the first two steps of the algorithm.

The deviance, number of parameters, and AIC for the preceding four models are displayed in Table 2. Whatever the chosen systematic part, we see that the gamma-Poisson model performs better than the independence negative-binomial model, thereby showing the need for modeling autocorrelation within unit.

Table 2
Deviance and AIC table

	Independence negative binomial			Gamma-Poisson		
Syst. part	Deviance	No. of parameters	AIC	Deviance	No. of parameters	AIC
Polynomial	553.7	6	565.7	546.2	8	562.2
Generalized logistic	556.5	5	566.5	548.8	7	562.8

The AIC provides no clear-cut choice for the systematic part. Therefore, it seems advisable to select the gamma-Poisson model with a generalized logistic growth curve for the systematic part because it is modeling the asymptotic behavior of the colony development for large time values.

The equations for the profiles of the gamma-Poisson model are given by, respectively,

$$\widehat{\mu}_{ij} = 0.6850 + 1097.4t_{ij} - 66.02t_{ij}^2 + 0.9843t_{ij}^3 + 0.0139t_{ij}^4$$

and

$$\widehat{\mu}_{ij} = \mathrm{e}^{6.304}\left[1 + (\mathrm{e}^{5.791} - 1)\mathrm{e}^{-0.0004t_{ij}\mathrm{e}^{7.600}}\right]^{-0.830}$$

for the polynomial and generalized logistic growth curves. They are both plotted on Figure 1. One can easily see that the polynomial model will predict an explosive number of individuals for large time values, whereas that number is estimated by $\mathrm{e}^{\widehat{\kappa_2}} \doteq 547$ for the logistic growth curve model.

Note that the serial association parameters for the preceding models are estimated by, respectively, $\widehat{\phi} = 0.2245$, $\widehat{\delta} = 0.000084$, $\widehat{\alpha} = 0.05937$ and $\widehat{\phi} = 0.1944$, $\widehat{\delta} = 0.000077$, $\widehat{\alpha} = 0.04837$ for the two best models.

5. Discussion

The gamma-Poisson model (Lambert, 1996) is extended to enable reduction of the mixing distribution into either a point or a vague prior. An extra parameter also enables to decrease the sensitivity of the model to extreme observations. These two modifications appear to be necessary when the merits of the gamma-Poisson and of the independence negative binomial model are compared using the Akaike criterion.

Note that the model described in the preceding sections can easily cope with sampling at irregularly spaced time points. Such situations might be caused by the design. For example, a regular sampling is very important at the beginning of the colony development (which undergo quick changes at the early stage), whereas its asymptotic behavior allows for more sparse observations after day 10. But coping with unequally spaced observation time may also be needed when missing completely at random data appear in an equally spaced sampling design.

Logistic growth curves (Nelder, 1961, 1962) are used to model the profile of the number of individuals in colonies of *Paramecium aurelium* in a count data context. That family of models is more realistic than a polynomial or any spline-based systematic part because each of the parameters has an interpretation as explained earlier. Moreover, it is more sensible than polynomials to model biological mechanisms of growth.

Covariates such as an indicator of the growth condition can easily be included as shown by Lindsey (1993, p. 133). They might affect the level of the asymptote or the growth rate of the colony at an early stage.

The parameters used by the full likelihood approach of this paper were estimated using the procedure OPTMUM in GAUSS, thereby enabling the use of nonlinear expressions for the mean.

ACKNOWLEDGEMENTS

I thank Jim Lindsey for his comments and the Camille Héla Foundation (University of Liège, Belgium) for its financial support.

RÉSUMÉ

Une version 'robuste' du modèle gamma-Poisson (Lambert, P., 1996, *Applied Statistics,* sous presse) pour des séries de comptages observés en des temps irrégulièrement espacés, est utilisée pour analyser la croissance de trois colonies isolées de paramécies (Diggle, P. J., 1990, *Time Series. A Biostatistical Introduction*). Les comptages relevés en des temps successifs dans une même série sont statistiquement dépendants. Une forme généralisée de la courbe de croissance logistique (Nelder, J. A., 1961, *Biometrika* **17,** 89–100; 1962, *Biometrics* **18,** 614–616) plus amplement dévelopée par Heitjan (1991, *Statistics in Medicine* **10,** 1077–1088; 1991, *Journal of the American Statistical Association* **86,** 891–898) et comprenant les courbes de croissance Mitscherlich, Gompertz, logistique et exponentielle, a été choisie pour modéliser les profils. Les comparaisons entre modèles (éventuellement non emboîtés) ont été réalisées avec l'aide du critère d'Akaike (Akaike, H., 1973, in *Second International Symposium on Inference Theory Petrov*).

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Inference Theory Petrov* B. N. and F. Csàki (eds), 267–281. Budapest: Akadémia Kiadó.

Diggle, P. J. (1990). *Times Series. A Biostatistical Introduction.* Oxford: Oxford University Press.

Fahrmeir, L. (1992). Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalized linear models. *Journal of the American Statistical Association* **87,** 501–509.

Harvey, A. C. and Fernandes, C. (1989). Time series models for count or qualitative observations. *Journal of Business and Economic Statistics* **7,** 407–423.

Heitjan, D. F. (1991a). Generalized Norton–Simon models of tumor growth. *Statistics in Medicine* **10,** 1075–1088.

Heitjan, D. F. (1991b). Nonlinear modeling of serial immunologic data: A case study. *Journal of the American Statistical Association* **86,** 891–898.

Jones, R. H. (1993). *Longitudinal Data with Serial Correlation: A State-Space Approach.* London: Chapman and Hall.

Jones, R. H. and Ackerson, L. M. (1990). Serial correlation in unequally spaced longitudinal data. *Biometrika* **77,** 721–731.

Jones, R. H. and Boadi-Boateng, F. (1991) Unequally spaced longitudinal data with AR(1) serial correlation. *Biometrics* **47,** 161–175.

Lambert, P. (1996). Modelling of repeated series of count data measured at unequally-spaced times. *Applied Statistics*, in press.

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

Lindsey, J. K. (1994). A pure likelihood inference derivation of the AIC or why likelihood ratio tests can be misleading. *Publications du Laboratoire de Statistique et Probabilités, Toulouse* **10,** 56–62.

Lindsey, J. K. and Lambert, P. (1995). Dynamic generalized linear models and repeated measurements. *Journal of Statistical Planning and Inference* **47,** 129–139.

Nelder, J. A. (1961). The fitting of a generalization of the logistic curve. *Biometrika* **17,** 89–100.

Nelder, J. A. (1962). An alternative form of a generalized logistic equation. *Biometrics* **18,** 614–616.

Ord, K., Fernandes, C. and Harvey, A. C. (1993). Time series models for multivariate series of count data. In *Developments in Times Series Analysis. In Honor of Maurice B. Priestley,* T. Subba Rao (ed), 295–309. London: Chapman and Hall.

Received November 1994; *revised March* 1995; *accepted April* 1995.