Sri Lankan Journal of Applied Statistics, Vol.9, 2008, p.159-175

Fitting Correlation Adjusted Generalized Linear Models to Clustered Dengue Data Measured Over Time

H L.C. PERERA AND M.R. SOORIYARACHCHI

Department of Statistics
University of Colombo, Sri Lanka

ABSTRACT

In this paper the authors address the issue of fitting generalized linear models (GLM's) in the presence of correlated observations, particularly when the data is clustered. In the usual GLM the responses obtained on each unit are considered independent. In this case the commonly used approach for the estimation of parameters is the method of maximum likelihood. However if correlation is present and is not taken into account then the standard errors of the parameter estimates will not be valid. One method of solution to this issue is estimation using Generalized Estimating Equations (GEE). In this paper the procedure involved in fitting GLM's with GEE method of estimation is discussed in detail and is illustrated using a set of data of dengue incidence in Sri Lanka in the years 2004 and 2005. The primary objective of this paper is to illustrate how generalized linear models can be fitted, in the presence of correlated data, by using generalized estimating equations. The secondary objective is to determine the factors effecting dengue incidence. This study showed that the dengue incidence can be adequately modeled using a GLM with negative binomial response. Patients within districts are believed to be more similar than patients in different districts and therefore a cluster effect is assumed within district. Also responses were believed to be correlated over time. This correlation structure is accommodated by using a autoregressive procedure of order one within the GEE framework. Rainfall and temperature in the current month and previous months up to a lag of two months were shown to effect incidence of dengue. In addition to these climatic variables dengue incidence also shows a changing pattern over time.

Key words: Generalized Linear Models, Generalized Estimating Equations, Correlated Data, Negative Binomial Distribution, Incidence

INTRODUCTION

Generalized linear models were formulated by Nelder and Wedderburn (1972) as a way of unifying statistical models with responses belonging to the exponential family. The generalized linear model (GLM) as explained by Dobson (2002) is a

flexible generalization of ordinary least squares regression. It relates the random distribution of the response variable of a study to the systematic linear predictor of the study through a function called the link function.

The commonly used approach for the estimation of parameters in the GLM is the method of maximum likelihood. The observations are assumed independent under the usual GLM. However when data are collected on the same unit across successive points in time, these repeated observations are correlated over time. If the correlation is not taken into account then the standard errors of the parameter estimates will not be valid and hypothesis testing results will be non replicable. One method of solution for this issue is estimation using Generalized Estimating Equations (GEE).

Generalized Estimating Equations (GEE) are methods of parameter estimation for correlated data. GEE was introduced by Liang and Zeger (1986) as a method of estimation of regression model parameters when dealing with correlated data. GEE methodology is a common choice when the outcome measure of interest is discrete (e.g. binary or count data, possibly from a binomial, poisson or negative binomial distribution) rather than continuous.

This paper illustrates the fitting of GLM's to correlated response data. The example taken is of dengue incidence in Sri Lanka in the years 2004 and 2005. The response variable, number of dengue cases, is given for 25 districts along with potential explanatory climatic variables, rainfall and temperature on a monthly basis. Patients within districts are believed to be more similar than patients in different districts and therefore a cluster effect is anticipated within district. Also responses are believed to be correlated over time.

The primary objective of this paper is to present in detail how generalized linear models can be fitted, in the presence of correlated data, by using generalized estimating equations. The secondary objective is to determine the factors effecting dengue incidence.

Section 2 explains the methodology involved in fitting GLM's using GEE method of estimation, section 3 consists of an example which applies the methods explained in section 2 to a set of dengue data. Finally, section 4 involves a discussion on the methods used and the results obtained.

METHODOLOGY

Generalized Linear Models (GLM's)

Nelder and Wedderburn (1972), McCullah and Nelder (1989) and Dobson (2002), among several other authors explain the development of GLM's in the presence of independent data, in great detail.

Correlation of Responses

When dealing with data that consist of measures that may be correlated within a cluster the correlation within responses must be accounted for. Otherwise incorrect inferences about the model coefficients can be made. (Because of incorrect estimation of the variances)

Ordinary Least Squares (OLS) regression models have been adopted for analysis of correlated responses when the dependent variable is normally distributed. But in conducting regression analysis of cluster-correlated binary or count dependent variables, one needs to use the quasi-likelihood method based on generalized linear models known as GEE's.

Theory of GEE's (Liang and Zeger, 1986)

Let Y_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots, K$ represent the j^{th} measurement on the i^{th} cluster. There are n_i measurements on cluster i and $\sum_{i=1}^{k} n_i$ total measurements.

Correlated data are modeled using the same link function and linear predictor setup (systematic component) as the independence case. The random component is described by the same variance function as in the independence case, but the covariance structure of the correlated measurements must also be modeled. Let the vector of measurements on the i^{th} cluster be $Y_i = [Y_{i1}, \ldots, Y_{i\,ni}]'$ with corresponding vector of means $\mu_i = [\mu_{i1}, \ldots\ldots, \mu_{ini}]^T$ and let V_i be the covariance matrix of Y_i. Let the vector of independent, or explanatory variables for the j^{th} measurement on the i^{th} cluster be $X_{ij} = [x_{ij1}, \ldots, x_{ijp}]'$

The Generalized Estimating Equation of Liang and Zeger(1986) for estimating the p×1 vector of regression parameters β is an extension of the independence estimating equation to correlated data and is given by

$$S(\beta) = \sum_{i=1}^{k} \frac{\partial \mu_i'}{\partial \beta} V_i^{-1} (Y_i - \mu_i(\beta)) = 0$$

$g(\mu_{ij}) = x_{ij}'\beta$ Here and g is the link function. The $p \times n_i$ matrix of partial derivatives of the mean with respect to the regression parameters for the i^{th} subject is given by

$$\frac{\partial \mu_i'}{\partial \beta} = \begin{bmatrix} \frac{x_{i11}}{g'(\mu_{i1})} & \cdots & \frac{x_{in_i1}}{g'(\mu_{in_i})} \\ \vdots & & \vdots \\ \frac{x_{i1p}}{g'(\mu_{i1})} & \cdots & \frac{x_{in_ip}}{g'(\mu_{in_i})} \end{bmatrix}$$

Working Correlation Matrix

$R_i(\alpha)$ denotes a $n_i \times n_i$ "working" correlation matrix that is fully specified by the vector of parameters α. The covariance matrix of Y_i is modeled as

$$V_i = \phi A_i^{\frac{1}{2}} W_i^{-\frac{1}{2}} R(\alpha) W_i^{-\frac{1}{2}} A_i^{\frac{1}{2}}$$

where A_i is an $n_i \times n_i$ diagonal matrix with $\upsilon(\mu_{ij})$ as the j^{th} diagonal element and W_i is an $n_i \times n_i$ diagonal matrix with w_{ij} as the j^{th} diagonal where w_{ij} is a weight (specified with the WEIGHT statement in SAS, PROC GENMOD. If there is no WEIGHT statement, $w_{ij} = 1$ for all i and j). If is the true correlation matrix of Y_i, then V_i is the true $R_i(\alpha)$ covariance matrix of Y_i. The working correlation matrix is usually unknown and must be estimated. It is estimated in the iterative fitting process using the current value of the parameter vector β to compute appropriate functions of the Pearson residual

$$e_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{\upsilon(\mu_{ij})/w_{ij}}}$$

If you specify the working correlation as $R_0 = I$, which is the identity matrix, the GEE reduces to the independence estimating equation. For data that are correlated within cluster over time, table 1 gives the structure of the working correlation supported by the GENMOD procedure in SAS and the estimator used to estimate the working correlations.

Table 1: Structure of working correlation for data correlated within cluster over time.

Working Correlation Structure		Estimator
Autoregressive AR(1)	$Corr(Y_{ij}, Y_{i,j+1})=\alpha$ for $t = 0,1,2, \ldots, n_i-j$	$\hat{\alpha} = \frac{1}{(K_1 - p)\phi} \sum_{i=1}^{K} \sum_{j \le n_i - 1} e_{ij} e_{i,j+1}$ $K_1 = \sum_{i=1}^{K} (n_i - 1)$

Dispersion Parameter

The dispersion parameter ϕ is estimated by

$$\hat{\phi} = \frac{1}{N-p} \sum_{i=1}^{k} \sum_{j=1}^{n_i} e_{ij}^2$$

Where $N = \sum_{i=1}^{k} n_i$ is the total number of measurements and p is the number of regression parameters. (The square root of $\hat{\phi}$ is reported by PROC GENMOD in SAS as the scale parameter in the "Analysis of GEE Parameter Estimates Model-Based Standard Error Estimates" output table).

Fitting Algorithm

The following is an algorithm for fitting the specified model using GEEs. Note that this is not in general a likelihood-based method of estimation, so that inferences based on likelihoods are not possible for GEE methods.

1. Compute an initial estimate of β with an ordinary generalized linear model assuming independence.
2. Compute the working correlations R based on the standardized residuals, the current β, and the assumed structure of R.
3. Compute an estimate of the covariance:

$$V_i = \phi A_i^{\frac{1}{2}} W_i^{-\frac{1}{2}} \hat{R}(\alpha) W_i^{-\frac{1}{2}} A_i^{\frac{1}{2}}$$

4. Update β:

$$\beta_{r+1} = \beta_r + \left[\sum_{i=1}^{K} \frac{\partial \mu_i'}{\partial \beta} V_i^{-1} \frac{\partial \mu_i}{\partial \beta}\right]^{-1} \left[\sum_{i=1}^{K} \frac{\partial \mu_i'}{\partial \beta} V_i^{-1} (Y_i - \mu_i)\right]$$

5. Iterate steps 2-4 until convergence

Parameter Estimate Covariances

The *model-based* estimator of $\mathrm{Cov}(\hat{\beta})$ is given by

$$\sum_m (\hat{\beta}) = I_0^{-1}$$

where $I_0 = \sum_{i=1}^{k} \frac{\partial \mu_i'}{\partial \beta} V_i^{-1} \frac{\partial \mu_i}{\partial \beta}$

This is the GEE equivalent of the inverse of the Fisher information matrix that is often used in generalized linear models as an estimator of the covariance estimate of the maximum likelihood estimator of β. It is a consistent estimator of the covariance matrix of $\hat{\beta}$ if the mean model and the working correlation matrix are correctly specified. The estimator

$$\sum_e = I_0^{-1} I_1 I_0^{-1}$$

is called the *empirical*, or *robust*, estimator of the covariance matrix of $\hat{\beta}$ where

$$I_1 = \sum_{i=1}^{k} \frac{\partial \mu_i'}{\partial \beta} V_i^{-1} \mathrm{Cov}(Y_i) V_i^{-1} \frac{\partial \mu_i}{\partial \beta}$$

It has the property of being a consistent estimator of the covariance matrix of $\hat{\beta}$, even if the working correlation matrix is misspecified, that is, if $\mathrm{Cov}(Y_i) \neq V_i$. In computing M, β and ϕ are replaced by estimates, and $\mathrm{Cov}(Y_i)$ is replaced by the estimate

$$(Y_i - \mu_i(\hat{\beta}))(Y_i - \mu_i(\hat{\beta}))'$$

Fitting the GLM using GEE

To fit Generalized Linear models using the GEE methodology four entities need to be specified. These are, namely,

- The distribution of the response variable (must belong to the exponential family)
- The link function
- The explanatory variables
- The covariance structure of the repeated measurements

Ballinger (2004) gives details on how to specify these entities. GEE's estimate model coefficients and standard errors with sampling distributions that are asymptotically normal. These estimates can be applied to test main effects and interactions and can be used to evaluate categorical or continuous independent variables. GEE estimates are the same as those produced by OLS regression when the dependent variable is normally distributed and no correlation within response is assumed.

GEE starts with maximum-likelihood estimation of regression parameters (β) and the variance is calculated using the link function, which is a transformation function that allows the dependent variable to be expressed as a vector of parameter estimates in the form of an additive model. The GEEs also use a variance function that is a transformation matrix with a value calculated from the observed mean that is used in calculating the variances of the parameters that permit nonconstant variances for values of the mean because they can depend on a specified function of the mean . The outcome produces both a matrix of the βs and a matrix with the inverse of the variance. If it is assumed that the data are correlated, the variances are multiplied against a working matrix of correlation coefficients that corrects for correlation within subjects or clusters. This matrix can be estimated by the GEE method in a form that matches the expected correlation structure within the subject or cluster.

The output from these equations is then used in starting the procedure all over again in an iteratively reweighted least squares procedure that involves minimizing the extent of change in the parameter estimates from a perfectly fitted regression model. This procedure is continued until convergence.

Model Selection

The process of selecting model terms and the appropriate correlation structure for GEE models is complicated by the correlation within subject. Because the observations are not independent of each other, the residuals are not independent, and therefore common likelihood based methods and other measures of model fit from ordinary linear regression need to be adjusted. According to Ballinger (2004) decisions about testing whether coefficients are equal to 0 are most commonly made using a Wald statistic. The Wald test statistic can be calculated by dividing the estimate of the parameter by its standard error. This has a standard normal distribution for large samples. It can be used to test the significance of individual parameters.

Goodness of fit of the model (Ten-Have and Chinchilli, 1998)

Residuals from GEE regression models should be checked for the presence of outliers that may seriously affect the results. Assessment of fit of the negative binomial model may proceed with analyses of Pearson residuals and evaluation of the generalized Pearson statistic,

$$\sum_{i=1}^{n} \frac{\left[Y_i - E(\hat{Y}_i)\right]^2}{\hat{var}(Y_i)}$$

where $\hat{E}(Y_i)$ and $\hat{var}(Y_i)$ are the estimated expectation and variance of Yi under a given model. The associated degrees of freedom are n-p, where p is the number of estimated parameters in the model.

Example

Description of Data

The response variable was chosen as the number of dengue cases recorded from a certain district, in a particular month of the year 2004 to 2005. Jaroensutasinee *et al.*, (2005) and Nakhapakorn and Tripathi (2005) suggest that the peak incident seasons of dengue are identified to be around the monsoon rain seasons i.e. June-August and December-February. In addition these two seasons correspond to the highest and lowest temperatures (respectively) within a year. These facts prompted in choosing monthly rainfall and temperature as the explanatory variables of this study. Also the lag effects of these two variables were taken into account since usually, places with stagnant pure water favorable for the dengue causing mosquito, *Aedes aegypti* to be

bred are frequent after a rainfall season. The linear, quadratic and cubic effect of time was incorporated by considering effect of the month on the response. Terms corresponding to t (linear time term), t^2 (quadratic time term) and t^3 (cubic time term) were considered.

Hence the following variables were used in the study: Number of cases recorded (cases), Year (y), Month of the year (mon), District (dist), Mean rainfall of the month (rf), Mean rainfall of the previous month (rfl1), Mean rainfall of two months before (rfl2), Mean temperature of the month (tmp), Mean temperature of the previous month (tmpl1), Mean temperature of two months before (tmpl2), Mean temperature of three months before (tmpl3), Time (t), Quadratic time term (t^2)=t*t , Cubic time term(t^3)=t*t*t.

A generalized linear model was fitted using these factors as covariates and keeping reported number of dengue cases as the response variable. Since the data are collected on the same districts across successive months in the year, these repeated observations are correlated over time. Thus in this situation GEE method of estimation is used for the estimation of parameters.

Specification of the model

Initially to model the Y_{ij} as a function of the explanatory variables, a generalized linear model (GLM) was fitted using GEE methodology, with a poisson distribution for the responses. The GEE methodology was used as data are collected on the same units (districts) across successive points in time (month within year). The poisson distribution was used as the response corresponds to counts. The characteristic link function for Poisson distribution i.e. the log link was used. Autoregressive correlation structure as suggested by Ballinger (2004) was used since the responses are correlated within cluster (district) over time (month).

Significant variables for the model were selected based on the Wald statistic by using the forward selection procedure and the most appropriate model was chosen. The chosen poisson model was highly over-dispersed and thus this required another distribution for the response to be pursued. As indicated by Ten-Have and Chinchilli (1998) when the poisson model is highly overdispersed and there are a large number of zero counts, the negative binomial which is described by Anscombe (1950) and Bliss and Owen (1958) is an alternative for modelling the data. Thus, a generalized linear model was refitted under GEE methodology for the data, with negative binomial distribution for the response, log link (the usual link function for

negative binomial distribution) and autoregressive correlation structure. The next step involves selecting the most appropriate negative binomial model.

Selecting the most appropriate model using Wald test

Step 1: The Null model

First the null model was fitted using *PROC GENMOD* procedure in *SAS*. The resulting Wald statistic was 7.94 corresponding to a p-value of less than 0.0001.

The null model is:

$$\log(\mu_{ij}) = \beta_0 \qquad (1)$$

The p value indicates that the intercept is highly significant.

Step 2: Fitting the Main Effects to the null model

The Wald test was used to select the most significant main effect. Initially the p-value was looked at and the variable giving the smallest p-value corresponding to the most significant variable was selected. In instances where the p-value has the same magnitude the value of the Wald statistic (Z) was used. Here the largest absolute value of Z corresponds to the most significant variable. Table 2 gives these results.

Table 2: Results of fitting each main effect to the model (1)

Term added	(Z) Wald test statistic	Difference in DF	P value
rf	3.62	1	0.0003
rfl1	7.16	1	<0.0001
rfl2	2.00	1	0.0456
tmp	-6.85	1	<0.0001
tmpl1	4.99	1	<0.0001
tmpl2	6.26	1	<0.0001
tmpl3	4.73	1	<0.0001
t	1.66	1	0.0965
t^2	-0.85	1	0.3931
t^3	0.15	1	0.8787

As shown in table 2, the variable tmp (temperature) has the greatest effect on the response as its Wald test statistic is highest and the p value is the most significant (i.e. the smallest p value). Therefore the variable tmp is added to the null model. The model selection is continued further with the model that contains the main effect tmp.

Now the selected model is;

$$\log(\mu_{ij}) = \beta_0 + \beta_1(\text{tmp} \qquad (2)$$

Proceeding in this way by adding one variable at a time until no more variables are significant at the 5% significance level we obtain that in addition to tmp several other variables, namely, tmpl2 (temperature at lag 2 months), rfl1 (rainfall at lag 1 months), tmpl1 (temperature at lag 1 months), rfl2 (rainfall at lag 2 months), t^3 (cubic effect of time), t^2 (quadratic effect of time) and rf (rainfall of current month) are also significant and enter the model in the order given. Thus the most appropriate model is

$$\log(\mu_{ij}) = \beta_0+\beta_1(\text{tmp})+\beta_2(\text{tmpl2})+\beta_3(\text{rfl1})+\beta_4(\text{tmpl1})+\beta_5(\text{rfl2})+\beta_6(t3)+\beta_7(t2)+\beta_8(\text{rf}) \qquad (3)$$

Model (3) contains 8 variables out of a total 10 variable considered.

When selecting the most appropriate model, apart from selecting a model that adequately describes the data, the interpretation of the model should also be simple. Therefore without further investigation of the interactions it is more appropriate to investigate the adequacy of the selected model.

Goodness of fit of the model

Assessment of fit of the negative binomial model was made by analyzing the Generalized Pearson residuals and using the Pearson's chi-square statistic. Figure 1 gives a plot of the generalized Pearson residuals versus the fitted linear predictor.

Figure 1 shows no specific pattern and the residuals are distributed symmetrically about zero apart from very few observations which are lying in right and left tails of the distribution, indicating a random distribution of residuals.

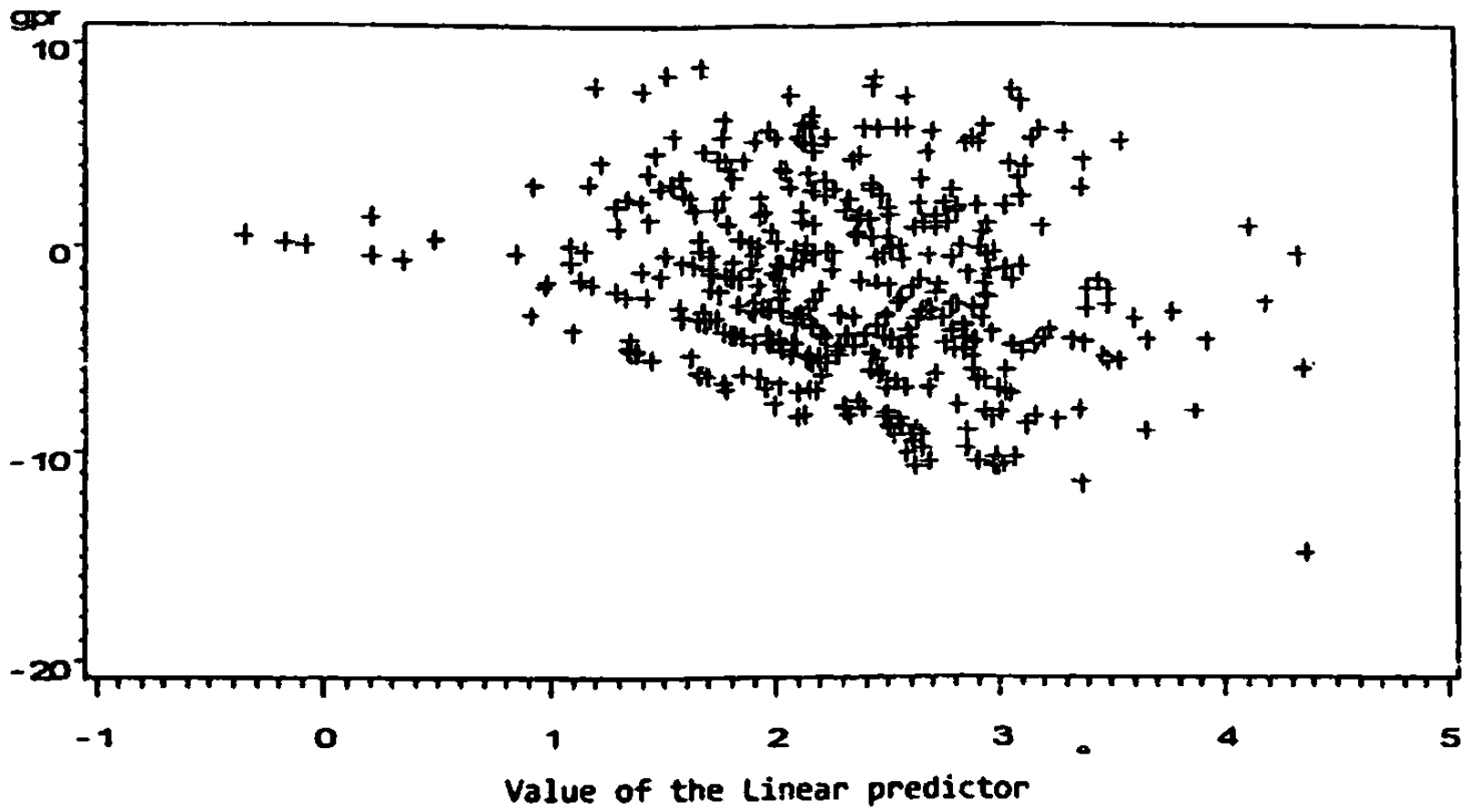

Figure 1:- Plot of Pearson residuals versus fitted values

The generalized Pearson's chi square statistic was 514.519 on 567 (=576-9) degrees of freedom resulting in a p-value of 0.944. Since the p-value is far greater than 0.05 the model (3) fits the data quite well. The residual analysis and the goodness of fit test indicate a well fitted model.

Parameter Estimates and Interpretation of the model

The model selected can be represented as

$$\log(\mu_{ij})= \beta_0+\beta_1(tmp)+\beta_2(tmpl2)+\beta_3(rfl1)+\beta_4(tmpl1)+\beta_5(rfl2)+\beta_6(t3)+\beta_7(t2)+\beta_8(rf) \quad (4)$$

$\log(\mu_{ij})$ gives the log of the expected number of dengue cases.

Parameter Estimates

After a model is fitted and its adequacy is established, the parameter estimates of the model should be interpreted. Table 3 gives the parameter estimates, standard errors and 95% confidence intervals for the estimates together with the corresponding Wald statistic and p-value.

Table 3: Parameter estimates of the selected Negative Binomial model

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr>|Z|
Intercept(β_0)	-2.9302	1.3076	-5.4932	-0.3673	-2.24	0.025
Tmp (β_1)	-0.2968	0.0374	-0.3701	-0.2234	-7.93	<.0001
tmpl2 (β_2)	0.2689	0.0320	0.2061	0.3316	8.40	<.0001
rfl1 (β_3)	0.0020	0.0002	0.0016	0.0025	8.97	<.0001
tmpl1 (β_4)	0.2010	0.0347	0.1330	0.2689	5.80	<.0001
rfl2 (β_5)	0.0016	0.0002	0.0012	0.0020	7.13	<.0001
t^3 (β_6)	-0.0017	0.0004	-0.0025	-0.0009	-4.23	<.0001
t^2 (β_7)	0.0151	0.0054	0.0046	0.0257	2.81	0.005
rf (β_8)	0.0005	0.0002	0.0001	0.0008	2.79	0.0053

The parameter estimates give the contribution of each variable to the log of the expected number of dengue cases recorded in each district of the country throughout the year. When the estimates of the parameters are considered it can be seen that all variables including the intercept are significant at 5% significant level.

Effect of temperature (tmp) on the response

The coefficient of tmp is negative indicating that increase of the temperature of the corresponding month will lead to decreasing of dengue cases. Suppose the temperature of a particular month increases by 1 unit and the expected number of dengue cases before and after this increment are μ_{1a} and μ_{2a} respectively.

$$\log(\mu_{1a}) = \beta_0+\beta_1(\mathbf{tmp})+\beta_2(tmpl2)+\beta_3(rfl1)+\beta_4(tmpl1)+\beta_5(rfl2)+\beta_6(t^3)+\beta_7(t^2)+\beta_8(rf) \quad (5)$$

$$\log(\mu_{2a}) = \beta_0+\beta_1(\mathbf{tmp+1})+\beta_2(tmpl2)+\beta_3(rfl1)+\beta_4(tmpl1)+\beta_5(rfl2)+\beta_6(t^3)+\beta_7(t^2)+\beta_8(rf) \quad (6)$$

$$(6)-(5) \Rightarrow \log(\mu_{2a}/\mu_{1a}) = \beta_1 = -0.2968$$

$$\Rightarrow \mu_{2a}/\mu_{1a} = \exp(\beta_1) = \exp(-0.2968) = 0.7432$$

$$\Rightarrow \mu_{2a} = 0.7432(\mu_{1a})$$

This result implies that expected number of dengue cases of a particular month decreases by a ratio of approximately 0.75, as a result of 1 unit increment in the temperature of that month. Using similar calculations it was found that,

1 the expected number of dengue cases of a particular month increases by a ratio of 1.2, as a result of 1 unit increment in the temperature of previous month.

2 the expected number of dengue cases of a particular month increases by a ratio of 1.3, as a result of 1 unit increment in the temperature of two months before.

3 the expected number of dengue cases of a particular month increases in a ratio of 1.0005, as a result of 1 unit increment in the rainfall of that month.

4 the expected number of dengue cases of a particular month increases in a ratio of 1.002, as a result of 1 unit increment in the rainfall of previous month.

5 the expected number of dengue cases of a particular month increases in a ratio of 1.0016, as a result of 1 unit increment in the rainfall of two months before.

Effect of time (t^2 and t^3) on the response

The variable t^2 indicates the quadratic effect of time whereas t^3 indicates the cubic effect of time. The coefficient of t^2 is positive while that of t^3 is negative. Suppose the time increases by 1 unit and the expected number of dengue cases before and after this increment are μ_{1g} and μ_{2g} respectively.

$$\log(\mu_{1g}) = \beta_0+\beta_1(tmp)+\beta_2(tmpl2)+\beta_3(rfl1)+\beta_4(tmpl1)+\beta_5(rfl2)+\beta_6(t^3)+\beta_7(t^2)+\beta_8(rf) \quad (7)$$

$$\log(\mu_{2g}) = \beta_0+\beta_1(tmp)+\beta_2(tmpl2)+\beta_3(rfl1)+\beta_4(tmpl1)+\beta_5(rfl2)+\beta_6(t+1)^3+\beta_7(t+1)^2+\beta_8(rf) \quad (8)$$

$$(8)-(7) \Rightarrow \log(\mu_{2g}/\mu_{1g}) = \beta_6[(t+1)^3 - t^3] + \beta_7[(t+1)^2 - t^2]$$

$$= \beta_6[t^3+3t^2+3t+1 - t^3] + \beta_7[t^2+2t+1 - t^2]$$

$$= \beta_6[3t^2+3t+1] + \beta_7[2t+1]$$

$$= 3\beta_6t^2 +(3\beta_6+2\beta_7)t+ (\beta_6+ \beta_7)$$

$$(\mu_{2g}/\mu_{1g}) = \exp[3\beta_6t^2 +(3\beta_6+2\beta_7)t+ (\beta_6+ \beta_7)]$$

Since (μ_{2g}/μ_{1g}) is not constant over time, in order to check the variation in (μ_{2g}/μ_{1g}) over time, a plot of (μ_{2g}/μ_{1g}) versus t (time) is plotted and given in figure 2.

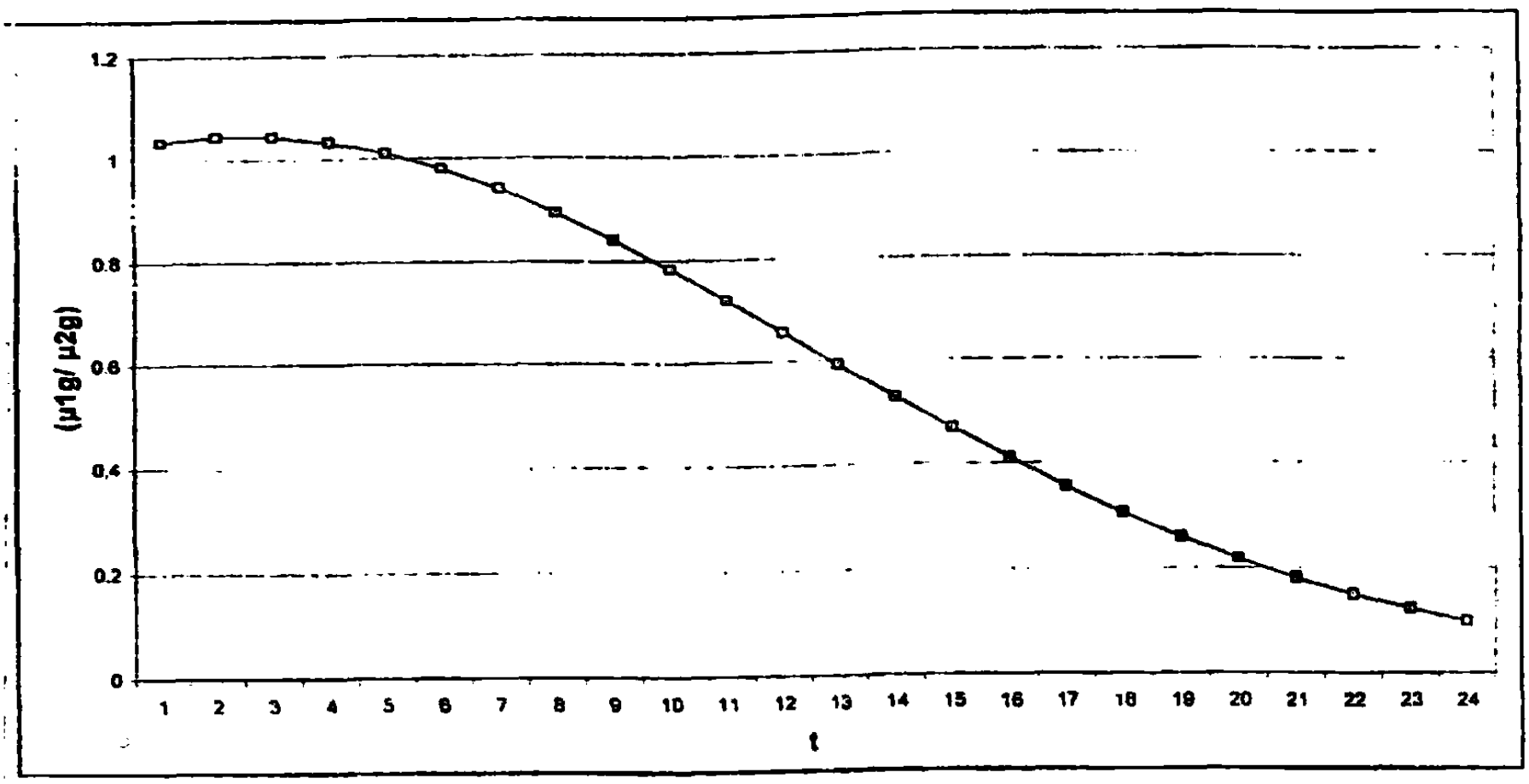

Figure 2: Plot of (μ_{2g}/ μ_{1g}) versus time

Figure 2 illustrates that expected number of dengue cases of a particular month increases in a ratio of over 1 within first 5 time points as a result of 1 unit increment in time. But after that in every time point expected number of dengue cases of a particular month decreases as a result of 1 unit increment in time. When observing the plot over the entire time scale it can be noticed that the ratio of dengue cases of the current month versus the previous month is a decreasing function of time. This indicates that there has been some control of dengue over the period of two years from 2004 to 2006.

DISCUSSION AND CONCLUSIONS

According to the main objective of this paper, the primary concern was to illustrate how the method of Generalized Linear Modeling can be implemented in the presence of correlated data. When correlation exists within data it is not acceptable to fit a GLM with ordinary maximum likelihood method of estimation. In this paper the authors initially explain the theory involved in fitting GLM's with GEE method of estimation. This involves detailed explanation of the procedure of estimation and the algorithm used in fitting the model, specification of the components of the model, model selecting procedure and goodness of fit technique.

Next an example is used to illustrate the procedure. This example is on a set of Dengue cases reported from Sri Lanka for the period of 2004-2005. The secondary objective was to reveal the factors which affect the incidence of dengue,

When considering the selection of variables for the model, according to the secondary objective, the response variable was chosen as the number of dengue cases recorded from a certain district, in a particular month of the year 2004 or 2005. Hence relevant to a single district, 24 time points were considered.

As Jaroensutasinee *et al.*, (2005), and Nakhapakorn and Tripathi (2005) suggest the importance of climatic variables in the prediction of dengue incidence, monthly rainfall and temperature were used as the potential explanatory variables. Also the lag effects of these two variables were taken into account since usually, places with stagnant pure water favorable for Aedes aegypti to be bred are frequent after a rainfall season. The time effect was incorporated by considering effect of the month on the response. Terms corresponding to t (linear time term), t^2 (quadratic time term) and t^3 (cubic time term) were considered.

The response, dengue incidence, was initially modeled as a function of the explanatory variables, using GEE methodology, with a Poisson distribution for the responses. However this model indicated the presence of over-dispersion and a large zero count for the response. In this case as suggested by Ten-Have and Chinchilli (1998), a negative binomial distribution was used instead of the Poisson distribution. The GEE methodology was used to accommodate the correlation in the data due to the cluster effect of the districts as well as that over time. The model specification consisted of taking the logarithm of the mean as the link function and the autoregressive correlation structure as the form of correlation of responses. The significant variables for the model were selected with the use of Wald statistics. Both, a plot of the Pearson residuals and a test based on the generalized Pearson statistic indicated that the model was adequately fitted.

Major findings suggested by the results are that dengue incidence is effected by both rainfall and temperature up to a lag of two months. Results also show that in general dengue cases have reduced over the time considered indicating possibly the success of dengue control programs (Hongisto, 2006).

ACKNOWLEDGEMENT

The authors would like to thank Dr. (Mrs.) P. Palihawadene and Mr. Premarathne of the Epidemiological Unit, Sri Lanka, for providing the dengue data.

REFERENCES

Anscombe, F.J. (1950). Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika*, 37: 358-382.

Ballinger, G.A. (2004). Using generalized estimating equations for longitudinal data analysis. *Organizational Research methods*, 7(2): 127-150.

Bliss, C.J. and A.R.G. Owen (1958). Negative binomial distribution and a common k. *Biometrika*. 45: 37-58.

Dobson, A. (2002). *Introduction to Generalized Linear Models*. Chapman & Hall, London.

Hongisto, K. (2006). *Dengue: Stopping a Potentially Deadly Threat in Sri Lanka.* Christian Children's Fund.

Jaroensutasinee, K., M. Jaroensutasinee and S. Promprou (2005) Climatic factors affecting dengue Haemorrhagic fever incidence in Southern Thailand. *Dengue Bulletin*, 29: 41-48.

Liang, K.Y. and S.L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73: 13-22.

McCullah, P. and J.A. Nelder (1989). *Generalized Linear Models, 2nd Edition.* Chapman and Hall, London.

Nakhapakorn, K. and N.K. Tripathi (2005). An information value based analysis of physical and climatic factors affecting dengue fever and dengue haemorrhagic fever incidence. *International Journal of Health Geographics*, 4: 13-35.

Nelder, J.A. and R.W.M. Wedderburn (1972). Generalized Linear Models. *Journal of the Royal Statistics Society. A*, 135: 370-384.

Ten-Have, T.R. and V.M. Chinchilli (1998). Two-stage negative binomial and over-dispersed poisson models for clustered developmental toxicity data with random cluster size. *Journal of Agricultural, Biological and Environmental Statistics*, 3: 75-98.