Volume: 10 Number: 03 Category: 02 Page: 49-63 ISSN No: 1303-6319 Cilt: 10 Sayı: 03 Kategori: 02 Sayfa: 49-63

IMPLEMENTATION OF REGRESSION MODELS FOR LONGITUDINAL COUNT DATA THROUGH SAS

Gül İNAN[*] **Özlem İLK** [**]

ABSTRACT

In this study, we firstly consider the marginal model and generalized linear mixed model classes for longitudinal count data and review the Log-Log-Gamma marginalized multilevel model, which combines the features of marginal models and generalized linear mixed models. Due to the special features of these models, implementation of them requires more special attention. As a consequence, this leads us to use SAS GENMOD procedure for the marginal model, SAS GLIMMIX procedure for the GLMM, and SAS NLMIXED procedure for the Log-Log-Gamma marginalized multilevel model. Since the latter model requires gamma distributed random effects, two different techniques, namely the probability integral transformation technique and likelihood reformulation technique , which are originally used for fitting Gamma Frailty models, are modified and adapted to fit Log-Log-Gamma marginalized multilevel model within the framework of Proc NLMIXED. Finally, we conclude the study with the discussion of the results obtained from the implementation of the models through popular epileptic seizures data.

Keywords: Epileptic seizure count, Gamma random effects, SAS GENMOD, SAS GLIMMIX, SAS NLMIXED.

1. INTRODUCTION

In longitudinal studies, measurements from the same subjects over a sequence of time periods are taken so that changes in measurements over time periods can be observed. In longitudinal count data (LCD), the response variable of the longitudinal dataset represents the counts of a total number of a defined event occurring in a given time interval. Examples from physiological research may include the number of epileptic seizures of each patient per two-weeks over an eight-week treatment period and the number of panic attacks for each patient in a week over a one-month psychological intervention program.

The analysis of longitudinal count data requires more special methods due to the longitudinal feature of measurements and counting process of responses. The most important feature of longitudinal data that motivates the statistical analysis is the association of measurements within a subject since the observations obtained from the same subject over several time periods are expected to be correlated. On the other hand, the statistical distribution of the counts is traditionally assumed to be Poisson distribution (Diggle et al., 2002) and it is well-known that the mean equals to the variance (equi-dispersion) for the Poisson distribution. However, when the variability of counts is greater than its expected value under the Poisson model, the phenomenon is

[*]Dr., Orta Doğu Teknik Üniversitesi, İstatistik Bölümü, Ankara, e-mail: ginan@metu.edu.tr
[**]Doç. Dr., Orta Doğu Teknik Üniversitesi, İstatistik Bölümü, Ankara, e-mail: oilk@metu.edu.tr

called overdispersion. More specifically, extra-Poisson variation occurs (Barron,1992). Although there are additional features that complicate the statistical analysis, these are the two that play a significant role in the estimation of regression parameters for the regression models developed for LCD.

In this sense, this paper aims to summarize the characteristics of the most commonly used general regression model classes, namely marginal models, random-effects models, and marginalized multilevel modelsto analyzelongitudinal count data and to show how these regression models are implemented through SAS, which is not well-documented in the literature, via the popular epileptic seizures example.On the other hand, the main contribution of this paper is that it provides the use of two different techniques to accommodate regression models with gamma distributed random effects for LCD, where non-gaussian random effects are not allowed within SAS framework.

The development of subsequent sections of this paper is organized as follows: Section 2 gives background information on regression model classes for LCD. Section 3 introduces the popular epileptic seizures example. Section 4 is devoted to the implementation of these regression models through the epileptic seizure example within SAS procedures. Section 5 discusses the results and Section 6 concludes the paper.

2. REGRESSION MODELS FOR LONGITUDINAL COUNT DATA

Diggle et al. (2002) classify the models for longitudinal data into three different regression model classes. These are: i) marginal models, ii) random-effects, and iii) transition models. In general, these three regression model classes view the association problem between the repeated measurements of a subject from different perspectives and this leads the models to differ in the interpretation of the regression parameters. In this paper, we restrict ourselves to the marginal and random-effects model classes and reintroduce the Log-Log-Gamma marginalized multilevel models (MMMs).

2.1 Marginal Models

Marginal models directly specify a regression model for the mean response, which depends only on covariates, using a log-link function**.** The mean responses, μ_{ij}^{C}, for the i^{th} subject and j^{th} time related to the covariates as follows:

$$log(\mu_{ij}^{C}) = X_{ij}\beta \qquad (1)$$

The within-subject association, the association between the repeated measurements of a subject, is modeled separately, possibly using additional association parameters**.** The regression parameters, β's, in equation (1) describe the effects of covariates on the population averaged mean response, as in cross-sectional analysis. Their interpretation is independent of specification of within-subject association model (Fitzmaurice and Molenberghs, 2008), which makes them more robust compared to the regression models that will be discussed later.

2.2 Random-Effects Models

The random-effects models assume that there is a natural heterogeneity between the subjects due to unmeasured covariates (Diggle et al., 2002). In this sense, regression parameters randomly varying from one subject to other subject are included into the regression modeling of the mean response. Contrary to the marginal models, GLMMs model the mean response and the within-subject association through a single equation and random effects are viewed as the potential source of within-subject association.

Among the random-effects models, generalized linear mixed models (GLMMs) are the most frequently used one for discrete repeated measurements (Molenberghs and Verbeke, 2005). In GLMMs, the model for the mean response depends both on covariates and random effects, which enter linearly into the linear predictor via a known link function. The simplest case of GLMMs is naturally a model with just a random intercept coefficient.

The formulation of a random-intercept model for LCD can be as follows:

i) Conditional Mean Model: $log(\mu_{ij}^{C}) = X_{ij}\beta + b_{0ij}$
ii) Random Intercept Distribution: $b_{0i} \sim MVN(0, C)$
iii) Conditional Response Distribution: $Y_{ij}^{C} = (Y_{ij}|b_{0ij}) \sim Poisson\ (\mu_{ij}^{C})$

Y_{ij}'s are assumed to be conditionally independent given subject-specific random intercepts,$b_{0i} = (b_{01i}, b_{02i}, \dots., b_{0in_i})'$ and to have Poisson distribution with conditional mean, μ_{ij}^{C}, depending on both fixed and random effects. The subject-specific random intercepts, $b_{0i} = (b_{01i}, b_{02i}, \dots, b_{0in_i})'$ are assumed to have a multivariate normal distribution with zero mean and a common within-subject covariance matrix, *C*.

One of the most important characteristics of GLMMs is that they have the ability to accommodate complex within-subject association structures for subject-specific random effects. Weiss (2005) lists a large number of covariance structures and detailed information on these covariance structure specifications, but among them, most commonly used ones are unstructured (UN), first order autoregressive (AR(1)), and compound symmetry (CS).

In GLMMs, the aim is to make inference on individual subjects rather than the population average; for that reason the fixed effects regression parameters, β's, in (i) describe the effects of covariates on an individual's mean response by controlling for the random-effects. However, interpretations being dependent on random effects and being sensitive to within-subject association specifications and robustness of estimates being dependent on the distribution of the random effects reflect the disadvantages of GLMMs (Heagerty and Zeger, 2000).

2.3 Log-Log-Gamma Marginalized Multilevel Model

Marginalized multilevel models are proposed by Heagerty and Zeger (2000). These models combine the features of marginal models and GLMMs with an aim to compensate the distinctions of these two models. While marginalized multilevel models take the interpretation and robustness of regression parameters from marginal models,

they take likelihood-based inference capabilities and flexible within-subject association specifications from GLMMs (Griswold and Zeger, 2004). Accordingly, Griswold and Zeger (2004) expand the marginalized multilevel model of Heagerty and Zeger (2000) for LCD and name this model as Log-Log-Gamma marginalized multilevel model (MMM).

The formulation of the Log-Log-Gamma MMM which assumes only a subject-specific intercept coefficient,b_{0i}, in the linear predictor, in addition to fixed effects, is as follows:

i) Marginal Mean Model: $log(\mu_{ij}^{M}) = X_{ij}\beta^{M}$

ii) Association Model: $log(\mu_{ij}^{C}) = \Delta_{ij} + b_{ij}$

iii) Random Effects Distribution:$g_{ij} \sim Gamma(1/\theta_1, \theta_2)$ where$b_{ij} = log(g_{ij})$

iv) Conditional Response Distribution: $Y_{ij}^{C} = (Y_{ij}|g_{ij}) \sim Poisson\ (\mu_{ij}^{C})$.

It defines a general linear model (GLM) for the marginal mean model in i) and a nonlinear mixed model (NLMM) for the within-subject association in ii).

Griswold and Zeger (2004) follow the same logic and assume a gamma distribution for subject-specific random effects and a Poisson distribution for the conditional response distribution, so that the marginal distribution of responses becomes negative-binomial distribution, which accommodates overdispersion well (Greenwood and Yule, 1920; Barron, 1992; Cameron and Trivedi, 1998; Jowaheer and Sutradhar, 2002). Contrary to GLMMs, subject-specific random effects in Log-Log-Gamma MMM follow a non-Gaussian distribution, that's Gamma distribution, and are allowed to enter nonlinearly into the model.

The log-link function and Poisson-gamma mixing distribution, together with the connection between marginal mean and conditional mean model, lead to $\Delta_{ij} = X_{ij}\beta^{M} - log\ (v_{ij})$where$v_{ij} = E(g_{ij}) = 1/\theta_1 \times \theta_2$(Griswold and Zeger, 2004). Hence, the conditional mean, μ_{ij}^{C}, can be written in terms of the marginal regression parameters, β^{M}, such that

$$\mu_{ij}^{C} = exp(\Delta_{ij} + b_{ij}) = exp(X_{ij}\beta^{M} - log(v_{ij}) + b_{ij}). \quad (2)$$

Since equation (2) includes the marginal regression parameters, β^{M}, the estimation of β^{M} can be performed by fitting the conditional model, μ_{ij}^{C}, via standard NLMM techniques. The regression parameters, β^{M}, describe the effects of covariates on the population averaged mean response, over the random effects.

3. EPILEPTIC SEIZURE COUNT DATA

The illustration of model fitting will be through an epileptic seizure count data, which is publicly availablein R package Mass (Venables and Ripley, 2002). We preferred this data set since it is the most commonly used one in the literature. This data comes from a randomized placebo-controlled clinical trial which was conducted by Leppik et al. (1985). 59 patients with simple or complex partial seizures were participated in the study and were randomized to receive either the antiepileptic drug progabide or a

placebo, as an adjuvant to the anti-epileptic standard chemotherapy. Before receiving treatment, the number of epileptic seizures of each patient over an eight-week period was recorded as baseline data. After treatment, the number of epileptic seizures of each patient per two-weeks over an eight-week treatment was also recorded at clinic visits. Apart from these, age information related to each patient was recorded as well. The question of interest is whether progabide has an effect in reducing the epileptic seizure counts or not.

The summary statistics for the epileptic seizure count data are displayed in Table 1. It is obvious that counts show overdispersion across visits within placebo group, progabide group, and complete data. When we do not take visits into account, the same case still continues, and counts exhibit high overdispersion in placebo group, progabide group, and complete data as an overall.

3.1 Covariates for Regression Models

To relate the covariates to the seizure counts, the covariates those listed in Thailand Vail (1990) are used. These are:

X_{1i}= logage=The natural logarithm of age in years, *log (Age)*,
X_{2i}= lgbsl=The natural logarithm of ¼ of the 8-week baseline counts,$log\ (Base/4)$,
X_{3i}= trt=Trt is a binary variable taking a value of 1 if progabide, 0 if placebo,
X_{4i}= v4=$Visit_4$ is a binary variable taking a value of 1 if visit number is 4, 0 otherwise,
X_{5i}= int=Interaction of Trt and $log\ (Base/4)$,

Here, β_3, which corresponds to the Trt variable, represent the parameter of interest for our research question.

Table 1. Summary statistics for epileptic seizure count data

	Placebo		Progabide		Complete	
	Mean	Variance / Mean	Mean	Variance / Mean	Mean	Variance / Mean
Visit 1	9.36	10.98	8.58	38.78	8.95	24.59
Visit 2	8.29	8.04	8.42	16.71	8.36	12.42
Visit 3	8.79	24.50	8.13	23.75	8.44	23.72
Visit 4	7.96	7.31	6.71	18.92	7.31	12.75
Overall	30.79	22.13	31.65	24.76	8.27	18.45

4. FITTING THE REGRESSION MODELS IN SAS

For model fitting of the regression models, SAS (version 9.2) is used.

4.1 Marginal Models

When the responses are discrete, i.e., binary or count, it is hard to estimate regression parameters of the marginal models by likelihood-based methods (Fitzmaurice and Molenberghs, 2008). That is because the complete joint distribution of longitudinal responses requires the specification of two-way associations between the responses and

in turn, building models for these associations that are consistent with the model for the mean response in an interpretable manner is difficult in the framework of marginal models (Lipsitz and Fitzmaurice, 2008).

When distributional assumption on repeated responses is avoided, an estimation method that is called generalized estimating equation (GEE) is considered. It is developed by Liang and Zeger (1986) by including additional parameters in the formulation of within-subject covariance matrix of responses. GEE provides as efficient estimates as maximum likelihood estimation (MLE), as well as consistent and asymptotically normal estimates provided that the mean response model is correctly specified. One disadvantage of GEE is that avoiding defining the complete joint distributions deprive us of using likelihood-based methods.

4.1.1 SAS GENMOD

SAS procedure that gives the opportunity to fit the GEE to repeated measures data is Proc GENMOD.

The marginal model equation for the epileptic seizure example can be given by

$$\mu = exp\left(\beta_0 + \beta_1 \times X_{log\,(Age)} + \beta_2 \times X_{log\,(\frac{Base}{4})} + \beta_3 \times X_{(Trt)} + \beta_4 \times X_{\left(Trt*log\left(\frac{Base}{4}\right)\right)} + \beta_5 \times X_{(Visit_4)}\right),$$

and the code related to our data and to our research question is given as follows:

```
procgenmoddata=seizure;
class id;
model count=logagelgbsltrtint v4 /dist=poissonlink=log scale=deviance;
repeated subject=id / type=UN;
run;
```

The model statement defines the relation between the response variable, count and the covariates logagelgbsltrtint v4, listed in Section 3. While dist option defines the distribution of counts, link option refers to the link function used in the model. On the other hand, scale=deviance enables the scale parameter to be fixed at 1 during estimation. Subjectthrough repeated statement identifies the subjects in the model and the variable identifying subjects should also be listed through the class statement. Finally, type refers to working correlation structure used in the model. SAS GENMOD allows user different working correlation structure types, such as unstructured, exchangeable and autoregressive AR (1).

4.2 Random-Intercept Model

When the interest is on the fixed effects regression parameters, β's, rather than random effects in the random-intercept model; the model fitting and inference on β's, requires the maximization of the likelihood of the data. This maximization is obtained by treating random intercepts, b_i's, as if they were nuisance parameters and by integrating over their distribution (Diggle et al., 2002). In other words, if the i^{th} subject's contribution to the likelihood of the data is defined as

$$L_i(\beta|\boldsymbol{Y_i}, \boldsymbol{b_i}) = \int_{b_i} \left[\left(\prod_{j=1}^{n_i} f_{ij}(y_{ij}|\beta, b_{ij}) \right) f(b_i|\theta) \right] db_i,$$

and then, the expression in equation (3)is expected to be maximized

$$L(\beta|\boldsymbol{Y}, \boldsymbol{b}) = \prod_{i=1}^{N} L_i(\beta|\boldsymbol{Y_i}, \boldsymbol{b_i})$$
$$= \prod_{i=1}^{N} \int_{b_i} \left[\left(\prod_{j=1}^{n_i} f_{ij}(y_{ij}|\beta, b_{ij}) \right) f(b_i|\theta) \right] db_i \quad (3)$$

where θ is the vector of parameters for the distribution of b_i.

In GLMMs, the distribution of random effects and high-dimensional integration of them together with a possibly nonlinear link function may cause computational difficulties in the evaluation of the likelihood and as consequence; closed-form solutions cannot be provided. In random-intercept models, being normal distribution not conjugate to Poisson distribution make the implementation of approximation harder.

Molenberghs and Verbeke (2005)divide the approaches toward the evaluation of the likelihood into three categories according to the frequency of usage and to the availability in statistical software. These are the approaches based on the approximation of i) the integrand, ii) data, and iii) integral itself. While Laplace–type approximations fall in the first category, penalized quasi-likelihood (PQL) and marginal quasi-likelihood (MQL) fall in the second category. The numerical integration methods such as adaptive and nonadaptive Gaussian quadrature fall in the latter category.

In this sense, SAS GLIMMIX procedure has the ability to fit the approximation and methods mentioned above.

4.2.1 SAS GLIMMIX

SAS GLIMMIX procedure is a built-in SAS procedure and is an appropriate choice for generalized linear mixed models, in which random effects are restricted to appear linearly in linear predictor. This procedure is especially recommended for models when the number of random effects per subject is large (Flom et al., 2006).

The random-intercept model equation for seizure data is given by

$$\mu = exp\Big(\beta_0 + \beta_1 \times X_{log\,(Age)} + \beta_2 \times X_{log\,(\frac{Base}{4})} + \beta_3 \times X_{(Trt)} + \beta_4 \times X_{\left(Trt*log\left(\frac{Base}{4}\right)\right)} + \beta_5 \times X_{(Visit_4)} + b\Big),$$

and $b \sim MVN_4(0, C)$ and C is assumed to be an unstructured within-subject covariance matrix.

SAS GLIMMIX code related to our data and to our research question can be given as follows:

```
procglimmix data=seizure MAXOPT=500 method=RSPL;
class id;
model count=logage lgbsl trt int v4 /dist=poisson link=log s;
random intercept /subject=id type=UN;
run;
```

Within the framework of Proc GLIMMIX, the random-intercept model is fitted by using PQL, based on REML for the linear mixed models. The option for PQL in **procglimmix** statement is the "method=RSPL", which is the default method. dist option through the model statement specifies the conditional distribution for the response variable given the random effects to come from any distribution in the exponential family. As in Proc GENMOD, link specifies the link function. interceptthrough random statement specifies a random intercept in the model. This procedure allows random effects to have only normal distribution and offer a straightforward fitting of a wide variety of within-subject covariance structures such as AR (1), CS and UN through type option.

4.3 The Log-Log-Gamma MMM

4.3.1 SAS NLMIXED

SAS NLMIXED procedure is a built-in SAS procedure and is preferred for the Log-Log-Gamma MM as in Griswold and Zeger (2004).

Proc NLMIXED is an appropriate choice for nonlinear mixed models, in which random effects are allowed to enter nonlinearly into the linear predictor of the model. It specifies the conditional distribution for the response variable given the random effects, either by standard distributions such as normal, binomial, and Poisson or by general distributions that can be coded using SAS statements. The only distribution available for random effects is normal distribution. The way of model specification in Proc NLMIXED has a high degree of flexibility, compared to other SAS procedures (Molenberghs and Verbeke, 2005). This advantage enables any non-normal distribution of interest for random effects to be implemented within the numerical integration techniques available in Proc NLMIXED via probability integral transformation(PIT) technique (Nelson et al., 2006) or likelihood reformulation (LR) technique (Liu and Yu, 2008). When the random effects are normally distributed, SAS NLMIXED procedure does not offer a straightforward option for the specification of any within-subject covariance structure. But, by the help of its flexibility, it is possible to allow the within-subject covariance matrix of the random effects to be, for instance, an AR(1), when specifying the mean and covariance components of the normal distribution (Molenberghs and Verbeke, 2005). Apart from these, Proc NLMIXED procedure requires the specification of initial values for all parameters in the model. Initial values for regression parameters can be obtained by the resulting parameter estimates after fitting a GLM in SAS.

In this sense, two different techniques, which Nelson et al. (2006) and Liu and Yu (2008) originally used for fitting Gamma Frailty models, are modified and adapted to fit

Log-Log-Gamma MMM by accommodating gamma distributed random effects within the framework of Proc NLMIXED.

1. PIT Technique by Nelson et al. (2006)

To accommodate gamma distributed random effects in Proc NLMIXED, we firstly use PIT technique proposed by Nelson et al. (2006). Similar to them, a_i is assumed to be a random effect from standard normal distribution, such that $a_i \sim N(0,1)$, and then by the use of PIT, it can be shown that $\Phi(a_i) = u_i \sim Unif(0,1)$ where $\Phi(.)$ is of the standard normal distribution. Again by the help of PIT, it can also be shown that $F_\theta(g_i) = u_i \sim Unif(0,1)$ where $F_\theta(.)$ is cumulative distribution function (CDF) of the gamma distribution of g_i, with $\theta = (1/\theta_1, \theta_2)$. For identifiably, θ_2 will be taken as equal to $\theta_1$on the forthcoming parts of the paper. Then it turns out that $g_i = F_\theta^{-1}(u_i) = F_\theta^{-1}(\Phi(a_i))$ has the gamma distribution of interest, where $F_\theta^{-1}(.)$is the inverse CDF of gamma distribution. Similarly, i^{th} subject's contribution to the likelihood of the data can be defined as in equation (4).

$$L_i(\beta|\boldsymbol{Y_i}, \boldsymbol{b_i}) = \int_{b_i} \left[\left(\prod_{j=1}^{n_i} f_{ij}(y_{ij}|\beta, b_{ij}) \right) f(b_i|\theta) \right] db_i, \tag{4}$$

where $b_i = log(g_i)$.

The expression in equation (5)**,** which is nowwritten in terms of random effects, a_i, is expected to be maximized such that

$$L(\beta|\boldsymbol{Y}, \boldsymbol{a}) = \prod_{i=1}^{N} L_i(\beta|\boldsymbol{Y_i}, \boldsymbol{a_i})$$
$$= \prod_{i=1}^{N} \int_{a_i} \left[\left(\prod_{j=1}^{n_i} f_{ij}\left(y_{ij}|\beta, F_\theta^{-1}(\Phi(a_i))\right) \right) \phi(a_i) \right] da_i \,, \tag{5}$$

where $\phi(.)$ is the standard normal distribution density function. Nelson et al. (2006) suggest that the likelihood in equation (5) can be approximated well by the Gaussian quadrature numerical integration technique. The approximation with Gaussian quadrature to integrals in equation (4) is achieved such that i^{th} subject's likelihood is approximated by a weighted sum

$$L_i(\beta|\boldsymbol{Y_i}, \boldsymbol{a_i}) = \int_{a_i} \left[\left(\prod_{j=1}^{n_i} f_{ij}\left(y_{ij}|\beta, F_\theta^{-1}(\Phi(a_i))\right) \right) \phi(a_i) \right] da_i$$
$$L_i(\beta|\boldsymbol{Y_i}, \boldsymbol{a_i}) \cong \sum_{q=1}^{Q} \left(\prod_{j=1}^{n_i} f_{ij}\left(y_{ij}|\beta, F_\theta^{-1}(\Phi(a_i))\right) \right) \phi(z_q) w_q,$$

and, thus, the likelihood in equation (5), which is expected to be maximized, turns out that

$$L(\beta|\boldsymbol{Y},\boldsymbol{a}) \cong \prod_{i=1}^{N}\sum_{q=1}^{Q}\prod_{j=1}^{n_i} f_{ij}\left(y_{ij}|\beta, F_{\theta}^{-1}(\Phi(a_i))\right)\phi(z_q)w_q,$$

where z_q is quadrature point and indexed by q = 1,...,Q, Q is the order of approximation, w_q is the standard Gauss-Hermite weight. Since the approximations will be more accurate as Q increases, we use Gaussian quadrature with 30 points like Griswold and Zeger (2004) and Nelson et al. (2006). The values of z_q and w_q can be obtained from tables in Abramowitz and Stegun(1972) (Table 25.10).

The Log-Log-Gamma MMM equation for seizure data is given by

$$\mu = exp\left(\beta_0 + \beta_1 \times X_{log\,(Age)} + \beta_2 \times X_{log\,(\frac{Base}{4})} + \beta_3 \times X_{(Trt)} + \beta_4 \times X_{\left(Trt*log\left(\frac{Base}{4}\right)\right)} + \beta_5 \times X_{(Visit_4)} + b\right),$$

and $b_{ij} = log(g_{ij}) \sim log - Gamma(1,1)$ or $e^{b_{ij}} = g_{ij} \sim Gamma(1,1)$.

SAS NLMIXED code by the help of PIT method that is related to our data and to our research question can be given as follows:

```
procnlmixed data=seizure noad fd qpoints=30;
PARMS theta1=1 beta0_m=-2.3492 beta1_m=0.7722 beta2_m=0.9582 beta3_m=-1.3299 beta4_m=-0.1565 beta5_m=0.5397;
eta_m=beta0_m + beta1_m*logage + beta2_m*lgbsl + beta3_m*trt + beta4_m*int + beta5_m*v4;
ui=CDF('Normal',ai);
if (ui >0.9999 ) then ui=0.9999;
gi1=quantile('GAMMA',ui,1/theta1,theta1);
v=1/theta1*theta1;
delta=eta_m-log(v);
eta_c=delta + log(gi1);
mu_c=exp(eta_c);
Model count ~ Poisson(mu_c);
Random ai ~ Normal(0,1) subject=id;
run;
```

noad in **procnlmixed** step refers to nonadaptive Gaussian quadrature. Finite difference approximation with fd is required for the derivative of CDF of normal distribution, that's CDF and the derivative of inverse CDF of gamma distribution, that's quantile. For that reason fd is there to specify that all derivatives to be computed using finite difference approximations. fd is equivalent to 100 as default and high fd values indicates better approximation. qpointsrefers to the number of quadrature points to be used during evaluation of integrals.PARMS statement allows to set the initial values for all unknown parameters in the model. The next eight SAS statements are used for defining Log-Log-Gamma MMM by PIT method. Model statementdefines the response variable and the form of the distribution of the conditional likelihood. Random statement declares the distribution of subject-specific random-intercept terms.

2. LR Technique by Liu and Yu (2008)

Another approach for accommodating gamma distributed random effects within the framework of the Proc NLMIXED is proposed by Liu and Yu (2008). This method aims to transform the formulation of likelihood that is conditional on non-normal random effects to a likelihood that is conditional on normal random effects in the framework of Gaussian quadrature. In this sense, they multiply and divide the likelihood in equation (6) by a standard normal density function, $\phi(.)$ such that

$$L_i(\beta|\boldsymbol{Y_i}, \boldsymbol{b_i}) = \int_{b_i} \left[\left(\prod_{j=1}^{n_i} f_{ij}(y_{ij}|\beta, b_{ij}) \right) f(b_i|\theta) \right] db_i \,, \tag{6}$$

$$L_i(\beta|\boldsymbol{Y_i}, \boldsymbol{b_i}) = \int_{b_i} \left[\left(\prod_{j=1}^{n_i} f_{ij}(y_{ij}|\beta, b_{ij}) \right) \frac{f(b_i|\theta)}{\phi(b_i)} \phi(b_i) \right] db_i \,,$$

$$L(\beta|\boldsymbol{Y_i}, \boldsymbol{b_i}) = \int_{b_i} \left[exp\left(log\left(\prod_{j=1}^{n_i} f_{ij}(y_{ij}|\beta, b_{ij}) \right) + log\big(f(b_i|\theta)\big) - log\big(\phi(b_i)\big) \right) \phi(b_i) \right] db_i \,,$$

$$= \int_{b_i} \left[exp\big(l_i^A + l_i^B - l_i^C\big) \phi(b_i) \right] db_i$$

where l_i^A is the conditional log-likelihood

$$l_i^A = log\left(\prod_{j=1}^{n_i} f_{ij}(y_{ij}|\beta, b_{ij}) \right) = log\left(\prod_{j=1}^{n_i} \left(\frac{e^{-\mu_{ij}^C} \mu_{ij}^{C\,Y_{ij}}}{y_{ij}!} \right) \right)$$

$$= -\sum_{j=1}^{n_i} exp\big(X_{ij}\beta^M - log\big(v_{ij}\big) + b_{ij}\big) + \sum_{j=1}^{n_i} y_{ij}\big(X_{ij}\beta^M - log\big(v_{ij}\big) + b_{ij}\big) - \sum_{j=1}^{n_i} log\big(y_{ij}!\big)$$

l_i^B is the log of the $log - Gamma \sim (1/\theta_1\,, \theta_2)$

$$l_i^B = log\big(f(b_i|\theta)\big) = -\frac{1}{\theta_1} log(\theta_1) - log\,\Gamma\left(\frac{1}{\theta_1}\right) + \frac{b_i}{\theta_1} - \frac{exp(b_i)}{\theta_1},$$

and l_i^C is the log of the standard normal distribution

$$l_i^C = log\big(\phi(b_i)\big) = -0.5 b_{ij}^2 + constant.$$

SAS NLMIXED code by the help of LR technique that is related to our data and to our research question can be given as follows:

```
procnlmixed data=seizure qpoints=30;
PARMS theta1=1 beta0_m=-2.3492 beta1_m=0.7722 beta2_m=0.9582 beta3_m=-
1.3299 beta4_m=-0.1565 beta5_m=0.5397;
eta_m=beta0_m + beta1_m*logage + beta2_m*lgbsl + beta3_m*trt + beta4_m*int +
beta5_m*v4;
v=1/theta1*theta1;
eta_c=eta_m-log(v)+ b;
mu_c=exp(eta_c);
expb=exp(b);
fc=fact(count);
loglik=-mu_c+count*eta_c-log(fc);
if lastid=1 then do;
lB=-((1/theta1)*log(theta1))-lgamma(1/theta1)+((1/theta1)*b)-((1/theta1)*expb);
lC=-1/2*(b**2);
loglik=loglik+lB-lC;
end;
Model count ~ general(loglik);
Random b ~ Normal(0,1) subject=id;
run;
```

Contrary to nonadaptive Gaussian quadrature**,**theadaptive Gaussian quadrature considers the shape of the likelihood when placing quadrature points and this result in better approximations (Liu and Yu,2008). For that reason, this technique prefers adaptive Gaussian quadrature contrary to PIT method which is the default option in **procnlmixed.** The next four SAS statements after the PARMS statement are there to specify the Log-Log-Gamma MMM. Similarly, Model statementshows the response variable and the form of the distribution of the conditional likelihood but this time through a general log-likelihood.

Contrary to PIT technique which requires the inverse CDF to have a closed form or to be available in SAS, LR technique requires that distribution function of the non-normal random effect to have a closed form or to be available in SAS. Further information on the description of the SAS NLMIXED and SAS GLIMMIX procedures and their options can be obtained from SAS (2000).

5. FINDINGS

Table 2 displays the regression parameter estimates and corresponding standard errors produced from the models and estimation methods mentioned above through the epileptic seizure data.

We find that results from four methods are similar except the estimates of regression parameter,β_4**.** Large differences are observed in this parameter between the regression models. It is found that the treatment effect has a statistically significant effect on the number of seizures count. As the negative sign on $\beta_3$indicates, the treatment reduces the seizure numbers.

Table 2. Results from the Marginal Model (ProcGenmod), Random-Intercept Model (Proc GLIMMIX), Log-Log-Gamma MMM by PIT (Proc NLMIXED) and Log-Log-Gamma MMM by LR (Proc NLMIXED)

Parameter	Marginal Model	Random-Intercept Model	Log-Log-Gamma MMM by PIT	Log-Log-Gamma MMM by LR
β_0	-2.5426 (0.9051)	-0.8776 (1.1217)	-1.1020 (0.5663)	-0.7594 (1.0863)
β_1	0.8417 (0.2608)	0.3558 (0.3259)	0.3556 (0.1680)	0.3378 (0.3195)
β_2	0.9455 (0.0931)	0.8780 (0.1369)	1.0774 (0.0823)	0.8926 (0.1273)
β_3	-1.4867 (0.4425)	-0.8671 (0.4139)	-1.4672 (0.2920)	-0.8173 (0.3826)
β_4	0.6019 (0.1789)	0.2984 (0.2096)	0.7044 (0.1013)	0.2971 (0.1914)
β_5	-0.1520 (0.0822)	-0.1565 (0.0544)	-0.1565 (0.0545)	-0.1565 (0.0545)

When we compare, the Log-Log-Gamma MMM by PIT and that by the LR method in terms of computational time, it is observed that LR method with adaptive Gaussian quadrature with 30 points option reduces the computational time considerably compared to PIT method with non-adaptive Gaussian quadrature 30 points option. While the estimation time takes approximately 2 seconds in LR technique, it takes about 22 seconds in PIT technique. This is due to that the LR technique does not need any finite difference approximation; hence it reduces implementation duration considerably.

6. CONCLUSION

This paper summarizes the marginal and random-effects model classes dealing with longitudinal count data in the literature and the implementation of longitudinal count data within SAS. One regression model class that is not mentioned in this paper is the transition models, but interested reader is kindly invited to read the Diggle et al. (2002). Diggle et al. (2002) reviews three different models and discusses the models with their pluses and minuses.

We especially focus on the Log-Log-Gamma marginalized multilevel model, which was developed by Griswold and Zeger (2004). This model is a likelihood-based model and offers a GLM for the mean response model, and a nonlinear mixed model for the within-subject association model. Separation of the model for mean response from that for within-subject association eases the interpretation of regression parameters of interest. Moreover, the Log-Log-Gamma MMM specifies a gamma distribution for the random effects which is conjugate to the Poisson distribution of conditional mean model. This is a great advantage over normally distributed random effects model since the Poisson-gamma mixture is able to remedy the overdispersion problem. As Nelson et al. (2006) stresses, non-normal random effects are taking progressive attention not only from longitudinal data analysis field, but also from different areas in statistics, and are more realistic than normally distributed random effects. However, non-normal random effects within the nonlinear mixed models suffer from the lack of computational implementation in the literature. In this sense, the main contribution of this paper is to show how a regression model with gamma distributed random effects, contrary to normally distributed random effects, can be handled within SAS, where non-Gaussian random effects are not allowed. We hope that the proposed algorithm would be helpful

for statisticians who work on models with non-Gaussian random effects and who would like to implement those models through user-specified algorithms within a standard software, i.e. SAS.

7. REFERENCES

Abramowitz, M., Stegun, I., 1972. Handbook of Mathematical Functions. Dover, New York.

Barron, D. N., 1992. The Analysis of count data: Overdispersion and Autocorrelation. Sociological Methodology, 22:179-220.

Cameron, A. C., Trivedi, P. K., 1998. Regression Analysis of Count Data. Econometric Society Monographs. Cambridge University Press, New York.

Diggle, P. J., Heagerty, P., Liang, K. Y., Zeger, S., 2002. Analysis of Longitudinal Data. Oxford University Press.

Fitzmaurice, G., Molenberghs, G., 2008. Advances in longitudinal data analysis: A historical perspective, in *Longitudinal Data Analysis: A Handbook of Modern Statistical Methods*, eds. G. Fitzmaurice, M. Davidian, G. Molenberghs, G. Verbeke, pp 3-27. Boca Raton, FL: Chapman & Hall/CRC Press, Florida.

Flom, P. L., McMahon, J. M., Pouget, E. R., 2006. Using PROC NLMIXED and PROC GLMMIX to analyze dyadic data with binary outcomes.Northeast SAS Users Group (NESUG) Proceedings, SAS Inc., Cary, NC.

Greenwood, M., Yule, G.U., 1920. An Inquiry into the Nature of Frequency Distributions Representative of Multiple Happenings with Particular Reference to the Occurrence of Multiple Attacks of Disease or of Repeated Accidents. Journal of the Royal Statistical Society, 83:255-279.

Griswold, M. E., Zeger, S. L., 2004. On Marginalized Multilevel Models and their Computation. The Johns Hopkins University, Department of Biostatistics Working Papers.

Heagerty, P. J., Zeger, S. L. 2000. Marginalized multilevel models and likelihood inference. Statistical Science, 15(1):1-26.

Jowaheer, V., Sutradhar, B. C., 2002. Analyzing longitudinal count data with overdispersion. Biometrika, 89(2):389-399.

Leppik, I. E., Dreifuss, F. E., Bowman-Cloyd, T., Santilli, N., Jacobs, M., Crosby, C., Cloyd, J., Stockman, J., Graves, N., Sutula, T., Welty, T., Vickery, J., Brundage, R., Gumnit, R., Gutierres, A., 1985. A double-blind crossover evaluation of progabide in partial seizures. Neurology, 35:285.

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

Lipsitz, S., Fitzmaurice, G., 2008. Generalized estimating equations for longitudinal data analysis. In *Longitudinal Data Analysis: A Handbook of Modern Statistical Methods*, eds. G. Fitzmaurice, M. Davidian, G. Molenberghs, G. Verbeke, pp 43-78. Boca Raton, FL: Chapman & Hall/CRC Press, Florida.

Liu, L., Yu, Z., 2008. A likelihood reformulation method in non-normal random effects models. Statistics in medicine, 27:3105-3124.

Molenberghs, G.,Verbeke, G., 2005. Models for Discrete Longitudinal Data. Springer, New York.

Nelson, K. P., Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J., Parzen, M., Strawderman, R., 2006. Use of the Probability Integral Transformation to Fit Nonlinear Mixed-Effects Models With Nonnormal Random Effects. Journal of Computational & Graphical Statistics, 15(1):39-57.

SAS/STAT User's Guide, Version 8, Chapter 46 (SAS Institute Inc.,Cary, NC, 2000).

Venables, W. N., Ripley, B. D. 2002. Modern Applied Statistics with S. Fourth Edition. Springer, New York.

Weiss, R. E., 2005. Modeling longitudinal data. Springer, New York.

UZUNLAMASINA KESİKLİ VERİLER İÇİN REGRESYON MODELLERİNİN SAS İLE UYGULANMASI

ÖZET

Bu çalışmada, öncelikle, uzunlamasına kesikli veri için marjinal model ve genelleştirilmiş lineer karma model sınıflarını ele alacağız ve sonra marjinal ve genelleştirilmiş lineer karma modellerinin özelliklerini birleştiren, Log-Log-Gamma marjinalleştirilmiş çok seviyeli modellerini yeniden gözden geçireceğiz. Bu modellerin bilgisayar ortamına aktarılması onların sahip olduğu birtakım özelliklerden dolayı, dikkat gerektirmektedir. Bu nedenden dolayı, bu durum marjinal modeller için SAS GENMOD, GLMM için SAS GLIMMIX ve Log-Log-Gamma marjinalleştirilmiş çok seviyeli modelleriiçin de SAS NLMIXED prosedürünü kullanmamıza öncülük etmektedir. Son model, gamma dağılımlı rassal etkiler içerdiğinden, ilk olarak Gamma Frailty modelleri için kullanılmış olan iki farklı yöntem, isim vermek gerekirse, olasılık integral dönüşümü ve olabilirlik yeniden formülasyonu yöntemleri değiştirilerek, Log-Log-Gamma marjinalleştirilmiş çok seviyeli modelleri için PROC NLMIXED prosedürü çerçevesinde uyarlanmıştır. Son olarak, çalışmamızı bu modellerin popüler epilepsi nöbet sayısı verisine uygulanmasından elde edilen sonuçları tartışarak bitirmekteyiz.

Anahtar Kelimeler: Epilepsi nöbet sayısı, Gamma dağılımlı rassal etkiler, SAS GENMOD, SAS GLIMMIX, SAS NLMIXED.