Vol. 2, No. 1, June 2018
ISSN: 2616-1303 | **Web:***www.cjpas.fulokoja.edu.ng*

Monte Carlo Studies on Conway-Maxwell Poisson Generalized Linear Mixed Effects Model for *Under-Dispersed* Count Data.

Shobanke Dolapo Abidemi,*Hussain Garba Dikko, O.E Asiribo , B.B Alhaji.
[1]Dept of Mathematical Sciences, Fed. University Lokoja Kogi State.
[2]Dept of Statistics, Ahmadu Bello University, Zaria Nigeria.
[3]Dept. of Statistics, University of Agriculture, Abeokuta Nigeria.
[4]Nigerian Defence Academy, Kaduna Nigeria.

Abstract.

Poisson regression is the traditional technique for handling count data. The assumption of equality of mean and variance which is an important property of the Poisson distribution makes the application of the distribution on count data highly restrictive since in reality count data do not always satisfy this assumption. The Generalized Poisson distribution and the Conway-Maxwell Poisson regression are some of the proposed remedies for handling under dispersed data. Our recent work on theoretical exposition of the re-parameterization and extension of the Conway-Maxwell-Poisson regression models to accommodate random effects appeared in the literature. This paper presents a simulation study to evaluate the performance of the re-parameterized Conway-Maxwell-Poisson Generalized Linear Mixed Effects Model (CMPGLMM) for handling the problem of under-dispersion in clustered data. The re-parameterization allows the response to be directly related to the regression coefficients via an approximation of the mean, thereby, leading to straightforward interpretation of the coefficients. The simulation result showed that the implementation is reliable and the CMPGLMM produced results that are better than the traditional Poisson and Negative-Binomial models which imply that the CMPGLMM is a better alternative for under-dispersed clustered count data.

Key words: *Under-dispersion, Clustered data, Poisson regression, Mixed-effect, Monte Carlo* Studies.

1. INTRODUCTION

Generalized linear mixed models (GLMMs) also known as multilevel generalized linear models (GLMs) are popular for multilevel data with units nested in clusters. GLMMs combine the properties of GLMs and linear mixed effects models (LMMs). As GLMs they have the ability to fit non-linear and non-Normal response data by using link functions and responses drawn from distributions in the exponential family. As mixed models they have the ability to include both fixed and random effects. Mixed-effects models represent the covariance structure related to the clustering of data by associating the common random effects to observations that have the same level of a clustering variable.

Poisson distribution using the mixed effect

***Corresponding Author**
How to cite this paper: Shobanke Dolapo Abidemi, Hussain Garba Dikko, O.E Asiribo , B.B Alhaji. (2018). Monte Carlo Studies on Conway-Maxwell Poisson Generalized Linear Mixed Effects Model for *Under-Dispersed* Count Data. *Confluence Journal of Pure and Applied Sciences (CJPAS), 2*(1), 147- 159.

framework is a traditional method for handling count data. One of such technique is the Poisson-Gamma distribution, though unsuitable for under-dispersed data McCullagh and Nelder (1997) also noted that the procedure is an unpopular option with problematic link. The Com-Poisson regression proposed by Sellers and Shmueli (2010) was recently extended by Dikko et al (2017) to accommodate random effects for handling clustered count data which are frequently encountered in observational or experimental studies.

Statistical methods for the analysis of cross-sectional count data where only one measurement is made for a variable of interest for each individual/observational unit in the study are well developed in literature. An important assumption for modelling cross-sectional data is that observations are independent of each other. Therefore, statistical methods for analyzing cross-sectional data cannot directly be used for analyzing longitudinal or clustered data. Clustered data can be defined as data in which the observations are grouped into disjoint classes called clusters according to some classification criterion (Pinheiro, 1994). These includes longitudinal data where individuals in a longitudinal setting are followed over a period of time and data collected at multiple time point for each individual (Wu, 2010). Here observations from each individual constitute a cluster. Mixed models were developed to handle clustered data and have attracted lots of interest in Statistics for decades. Observations in the same cluster usually cannot be considered independent therefore mixed effects models constitute a convenient tool for modelling cluster dependence.

Dikko et al (2017) combined the ability of the GLMMs to account for correlation within clustered data and the flexibility of the COM-Poisson distribution in handling any dispersion level in count data to propose a COM-Poisson GLMM (CMPGLMM). In this paper, we present a simulation study to evaluate the performance of the CMPGLMM alongside the Poisson and Negative-Binomial GLMM in the presence of under-dispersion.

The rest of the paper is organized as follows. Section 2 provides an overview of count distributions and regression models; section 3 gives some details on the re-parameterization used by Dikko et al (2017) as well as some information on implementation; section 4 consists of the simulation setting, results and discussion while section 5 provides concluding remarks.

2. COUNT MODELS

2.1 Poisson Regression

The Poisson distribution characterizes the probability of observing any discrete number of events given an underlying mean count of events, assuming that the timing of the events is random and independent (Le, 2003).

The Poisson regression model is a model where the mean of the distribution is a function of the explanatory variables, with the defining characteristic that the conditional mean of the outcome

explanatory variables, with the defining characteristic that the conditional mean of the outcome is equal to the conditional variance (Algamal, 2012; Algamal and Lee, 2015).

In Poisson regression model, the number of events y_i has a Poisson distribution with a conditional mean that depends on individual characteristics according to the structural model;

$$P(Y_i) = \frac{e^{-\theta_i}\theta_i^{y_i}}{y_i!};\ y_i = 0,1\ldots;\ i = 1,2\ldots,n \quad (1)$$

and the conditional mean parameter $\theta_i = exp\left(X_i^T\beta\right)$, where, $\beta^T = \left[\beta_1,\beta_2\ldots\beta_p\right]$denotes a $1 \times p$ vector of regression parameters and X_i a$p \times 1$ vector, $p = k + 1$.

The interpretation of each coefficient depends on whether the corresponding covariate is categorical or continuous. If the covariates are continuous then $exp\left(\beta_j\right)$ represents a multiplicative effect of the X_j on the expected mean (Liao, 1994).

There are two main approaches for interpreting coefficients in regression models(Long, 1997). The first approach examines the changes in the conditional mean for a unit change in a single predictor via the additive or the multiplicative model. The second approach used in non-linear regression models is to examine

Vol. 2, No. 1, June 2018
ISSN: 2616-1303 | **Web:***www.cjpas.fulokoja.edu.ng*

Shmueli *et al*.(2005)used an asymptotic expression to derive the approximation for Z with the mean and variance given as:

$$E(Y) \approx \lambda^{\frac{1}{\nu}} + {}^{1}/_{2\nu} - {}^{1}/_{2} \quad (3)$$

$$Var(Y) \approx \frac{1}{\nu}\lambda^{\frac{1}{\nu}} \quad (4)$$

3. A NEW PARAMETERIZATION OF THE COM-POISSON DISTRIBUTION

Let $\omega = \lambda^{\frac{1}{\nu}} + \frac{1}{2\nu} - \frac{1}{2}$ which is the approximated mean of the distribution. Guikema and Coffelt (2008) expressed λ in terms of α, that is, $\lambda = \alpha^{\nu}$ and then model the response via

$$\alpha_i = \exp(X_i^T \beta)$$

Here, we express λ in terms of ω:

$$\lambda = \left(\omega - \frac{1}{2\nu} + \frac{1}{2}\right)^{\nu}.$$

The relationship between Y_i and X_i is modelled via

$$E(Y_i) \approx \omega_i = \exp(X_i^T \beta)$$

The Com-Poisson PMF under our re-parameterization is given as

$$P(y_i | \omega_i, \nu) = \frac{\left(\left(\omega_i - \frac{1}{2\nu} + \frac{1}{2}\right)^{y_i}\right)^{\nu}}{(y_i!)^{\nu} z(\omega_i, \nu)} \quad (5)$$

Vol. 2, No. 1, June 2018
ISSN: 2616-1303 | **Web:***www.cjpas.fulokoja.edu.ng*

Where $Z(\omega_i, \nu) = \sum_{h=0}^{\infty} \left(\frac{\left(\omega_i - \frac{1}{2\nu} + \frac{1}{2}\right)^h}{h!} \right)^{\nu}$.

Based on the re-parameterized Com-Poisson distribution given above, we present the formulation of our mixed effect model in the next subsection.

Let y_{ij} denote the jth response for the ith cluster, $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$. For each i, conditional on random effect b_i, the $y_{ij}, j = 1, \ldots, n_i$ are assumed to be independent and follow a Com-Poisson (CMP) distribution where the probability mass function of the CMP distribution using our proposed re-parameterization is

$$P(y_{ij} | b_i, \omega_{ij}, \nu) = \frac{\left(\left(\omega_{ij} - \frac{1}{2\nu} + \frac{1}{2}\right)^{y_{ij}}\right)^{\nu}}{(y_{ij}!)^{\nu} Z\left(\left(\omega_{ij} - \frac{1}{2\nu} + \frac{1}{2}\right), \nu\right)} \quad (6)$$

where,

$Z(\omega_{ij}, \nu) = \sum_{h=0}^{\infty} \left(\frac{\left(\omega_{ij} - \frac{1}{2\nu} + \frac{1}{2}\right)^h}{h!} \right)^{\nu}$ and

$E[Y_{ij}] \approx \omega_{ij}$ and $Var[Y_{ij}] = \frac{\omega_{ij} - \frac{1}{2\nu} + \frac{1}{2}}{\nu}$ where $\omega_{ij} > 0, \nu \geq 0, i = 1, 2, \ldots, N$; $y_{ij} = 0, 1, 2, \ldots$

The heirechical representation of our CMPMM formulation is,

$$y_{ij} | b_i \sim independent\ CMP(\omega_{ij}, \nu) \quad (7)$$

$$b_i \sim i.i.d\ N(0, \sigma^2)$$

and

Vol. 2, No. 1, June 2018
ISSN: 2616-1303 | **Web:***www.cjpas.fulokoja.edu.ng*

$$g\left(E\left(Y_{ij}|\, b_i\right)\right) = \log\omega_{ij} =$$

$$X_{ij}^T\beta + b_i \qquad (8)$$

and

$$\omega_{ij} = \exp\left(X_{ij}^T\,\beta + b_i\right). \qquad (9)$$

Details of the procedure for obtaining estimates of the parameters in the CMPGLMM can be found in Dikko *et al* (2017).

3.1 Implementation

The Conway-Maxwell Poisson Generalized Linear Mixed effect Model (CMPGLMM) has been implemented using R (R Core Team, 2017). To maximize the likelihood functions, we employ the Bound Optimization BY Quadratic Approximation (BOBYQA) Powell, (2009) algorithm which performs derivative-free bound-constrained optimization using an iteratively constructed quadratic approximation for the likelihood function. The algorithm is very robust for optimizing functions with many parameters. It uses a trust region method that forms quadratic models by interpolation. The algorithm optionally allows constraints to be placed on the parameters. For more details on the algorithm see (Powell, 2009). The BOBYQA algorithm adopted for this work is the one implanted by the `nloptr` R package version 1.0.4 (Johnson, 2017) which implements an R interface to NLopt. NLopt is a free/open-source library for nonlinear optimization which provides a common platform for a number of different free optimization routines as well as original implementations of various other algorithms.

Vol. 2, No. 1, June 2018
ISSN: 2616-1303 | **Web:***www.cjpas.fulokoja.edu.ng*

Various R functions were written to carry out specific tasks. For example, the functionCOMP_Z.Q and likfncompute $z(\theta_{ij}, v)$, $Q(b_i)$ and $\ell_p(\beta, \hat{v}, \hat{\sigma})$ (the profiled loglikelihood for β respectively). The main R function that is called to fit a CMPGLMM given a dataset is cmpfitme. Calling the function will make calls to various necessary functions and return the estimated coefficients, random effects variance as well as standard errors. The function allows the response variable, predictors as well as the clustering variables to be specified. An example of the function usage is

```
Cmpfitme(No_casualties~month+Age+Gender+Cause+Type_Acci
dent+Nature_road+(1|location)+(1|Type_Vehicle),data=cda
t)
```

In the above example, No_casualties is the count response variable, month, Age, Gender, Cause,Type_AccidentandNature_road are the predictors while location and Type_Vehicle are the clustering variables which will constitute random effects terms.

4. SIMULATION STUDIES

The performance of the CMPGLMM for estimation at various sample sizes, dispersion level is examined through empirical simulations vis-à-vis other clustered count modelling methods such as the Poisson GLMM (PGLMM) and the negative binomial GLMM (NBGLMM). All simulations and computations were carried out using R(R Core Team, 2017).

4.1 Simulation Setting

The true underlying model from which we simulate data is a model with one clustering variable and is given by

$$E(y_{ij}|X_{1ij}, X_{2ij}, b_i) = \theta_{ij}$$
$$= exp(\beta_0 + X_{1ij}\beta_1 + X_{2ij}\beta_2 + b_i), \quad (10)$$
$$b_i \sim N(0, \sigma)$$

$i = 1, \ldots, m$, $j = 1, \ldots, n_i$. The parameters of the model were set as follows: $\beta_0 = 0.2$, $\beta_1 = -2$, $\beta_2 = 0.3$ and $\sigma = 1$. The number of clusters was varied as $m \in \{5, 10\}$ and the number of observations per cluster was set as $n_i \in \{5, 10\}$. Hence, the sample size setting considered are: $m = 5$, $n_i = 5$ (total number of observations $n = 25$); $m = 5$, $n_i = 10$ (total number of observations $n = 50$); $m = 10$, $n_i = 10$ (total number of observations $n = 100$); Furthermore, the predictors were generated as follows: $X_1 \sim N(0,1), X_2 \sim Unif(0,2)$.

The under-dispersed distribution considered is the double Poisson (DPOIS) distribution (Efron, 1986; Ridout and Besbeas, 2004). The under-dispersed responses were simulated such that $y_{ij} \sim DPOIS(\theta_{ij}, 0.3)$.

This model has three fixed effects parameters (β_0, β_1 and β_2), adding the random effects variance parameter σ^2 makes the total number of parameters to be four. It is important to note that there is only one random effects term in the model under consideration, therefore m random effects will be estimated for each case.

Estimates of the Poisson GLMM (PGLMM) and the negative binomial GLMM (NBGLMM) were obtained using the algorithms implemented in the lme4 R package while the CMPGLMM estimates were obtained using our own R implementation. The performances of the methods are evaluated over 100 replications of each setting discussed above. The evaluation criteria are: average estimation error (AE_j) defined as $E(|\hat{\beta}_j - \beta_j|) = \frac{\Sigma|\hat{\beta}_j - \beta_j|}{100}$; mean-squared errors of

Vol. 2, No. 1, June 2018
ISSN: 2616-1303 | **Web:***www.cjpas.fulokoja.edu.ng*

estimates (MSE_{β_j}) defined as $E\left(\left[\hat{\beta}_j - \beta_j\right]^2\right) = \frac{\sum(\hat{\beta}_j - \beta_j)^2}{100}, j = 0,1,2$. Similarly, estimation of σ is also evaluated.

4.2 RESULTS/DISCUSSION

The results of the application of the techniques and simulation are presented in Table 1. Only the estimates of the major parameters (fixed effects and random effects standard deviation) are reported here.

The simulation results show that the CMPGLMM performed better and yielded better results than the PGLMM and NBGLMM when the correlated count data are under-dispersed. The simulation results also show that the estimate of the dispersion parameter $\hat{\nu}$ of the CMPGLMM varies according to the nature of dispersion exhibited by the count data. For example, the average estimates $\hat{\nu}$ for $m = 10$ (10 clusters) and $n_i = 10$ (10 observations per cluster) under-dispersion is 3.807. This implies that during modelling of clustered count data, using the CMPGLMM the method detects the type of dispersion and fits the corresponding model. Furthermore, the average estimated dispersion parameter $\hat{\nu}$ of the CMPGLMM is 4.009 for small sample size setting and 4.14 at the other sample size settings showing that the response data are highly under-dispersed. The CMPGLMM produced the lowest estimation and mean square errors for all parameters at all the different sample size settings. Also, the CMPGLMM produced the lowest errors for σ at all the sample size settings $(m = 5, n_i = 5)$, $(m = 5, n_i = 10)$ and $(m = 10, n_i = 10)$.

Table 1:*Average estimation errors (**AE**) and mean squared errors of estimation (**MSE**) for underdispersion based on 100 replications over three different sample size settings.*

Sample Size setting	*Method*	**Avergae Estimation Error**				**MSE**			
		β_0	β_1	β_2	σ	β_0	β_1	β_2	σ
$m = 5, n_i = 5$	***PGLMM***	0.858	0.061	0.083	0.394	50.114	0.374	0.169	0.377
	NBGLMM	0.858	0.061	0.083	0.394	50.113	0.374	0.169	0.377
	CMPGLMM ($\bar{\hat{\nu}} = 4.009$)	0.768	0.056	0.078	0.214	50.011	0.371	0.163	0.309
$m = 5, n_i = 10$	***PGLMM***	2.732	0.185	0.152	0.741	12.400	0.052	0.037	0.732
	NBGLMM	2.732	0.185	0.152	0.741	12.400	0.052	0.037	0.732
	CMPGLMM ($\bar{\hat{\nu}} = 4.14$)	2.639	0.136	0.146	0.207	11.495	0.028	0.035	0.066
$m = 10, n_i = 10$	***PGLMM***	1.862	0.087	0.105	0.499	6.812	0.012	0.021	0.313
	NBGLMM	1.861	0.087	0.105	0.499	6.812	0.012	0.021	0.313
	CMPGLMM ($\bar{\hat{\nu}} = 3.807$)	1.855	0.074	0.105	0.261	6.718	0.009	0.021	0.107

5. CONCLUSION

The implementation of the CMPGMM has been discussed. The performance of the Com-Poisson Generalized Linear Mixed Effects Model (CMPGLMM) has been evaluated compared to the Poisson and Negative Binomial linear mixed effects models (PGLMM and NBGLMM respectively) via Monte Carlo studies. The simulation result shows that our implementation is reliable. The implementation allows both the fixed effects and random effects parameters to be estimated at a relatively good computational cost. Also, the implementation allows the dispersion parameter ν to be estimated from which one can deduce the type of dispersion.

The results from the simulation show that CMPGLMM produced the best results among the three methods used at different sample size settings, i.e. the model outperform the PGLMM and the NBGLMM this is obviously due to presence of

under-dispersion in the response. The result here shows the versatility of CMPGLMM in handling under-dispersion in clustered count data.

REFERENCES

Algamal, Z. Y. (2012). Diagnostic in Poisson regression models. Electronic Journal of Applied Statistical Analysis, 5(2):178{186.

Algamal, Z. Y. and Lee, M. H. (2015). Adjusted adaptive lasso in high-dimensional Poisson regression model. Modern Applied Science, 9(4):170{177.

Dikko, H.G, Asiribo,O.E, Alhaji, B.B and Shobanke, D.A (2017) Modelling under dispersion in cluster count data using the modified Conway-Maxwell-Poisson regression. *Confluence Journal of Pure and Applied Sciences.* Vol 1 no1. Nov. 2017. ISSN-2616-1303. 105-130

Efron, B. (1986). Double Exponential Families and Their Use in Generalized Linear Regression. Journal of the American Statistical Association, 81:709-721.

Famoye, F. (1993).Restricted Generalized Poisson Regression Model. Communications in Statistics - Theory and Methods, 22(5):1335–1354.

Guikema, S. D., and Coffelt, J. P. (2008).A flexible count data regression model for risk analysis.*Risk Analysis*, **28**: 213-223.

Johnson, S. G. (2017). The NLopt nonlinear-optimization package, http://ab-initio.mit.edu/nlopt

Le, T. C. (2003).*Introductory Biostatistics*. John Wiley & Sons, Inc., New Jersey.

Long, J. S. (1997). Regression Models for Categorical and Limited Dependent Variables.Sage, London.

Liao, T. F. (1994).*Interpreting Probability Models: Logit, Probit, and Other Generalized Linear Models. Sage University Paper Series on Quantitative Applications in the Social Sciences*, 07-101, Sage, Newbury Park, CA.

McCullagh, P. and Nelder, J. A. (1997).*Generalized Linear Models, 2nd edition*.Chapman & Hall/CRC.

Penston MJ, Millar CP, Zuur AF, Davis IM (2008) Spatial and temporal distribution of Lepeophtheirus salmonis (Krøyer) larvae in a sea loch containing Atlantic salmon. Journal of Fish Diseases 31:361–371

Pinheiro, J.C. (1994), "Topics in Mixed Effects Models," PhD Thesis, University of Wisconsin-Madison.

Powell. M. J. D. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives.*Department of Applied Mathematics and Theoretical Physics, Cambridge England, technical reportNA2009/06*.

R Core Team (2017).*R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.URL https://www.R-project.org/.

Ridout, M.S & Besbeas, P. (2004).An empirical model for underdispersed count data. Statistical Modelling, 4(1): 77-89.

Sellers, K. and Shmueli G. A. (2010) Flexible Regression Model for Count Data. *Annals of Applied Statistics*,**4**:943–961. DOI: 10.1214/09-AOAS306.

Vol. 2, No. 1, June 2018
ISSN: 2616-1303 | **Web:***www.cjpas.fulokoja.edu.ng*

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., and Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution. *Applied Statistics*, 54:127–142.

Wu, L. (2010). Mixed Effects Models for Complex Data. Chapman & Hall/CRC, New York.