Adequacy of H-Likelihood Estimation Method for Unbalanced Clustered Counting Data Models.

Intesar N. El-Saeiti[1], Khalil Mostafa ALsawi[2] and Gebriel M. Shamia[3]

[(1)] *Assistant Professor, Statistics Department, Faculty of Science, University of Benghazi*
[(2)] *A teacher at a secondary school, Benghazi-Libya*
[(3)] *Professor, Statistics Department, Faculty of Science, University of Benghazi*

Correspondence to Intesar N. El-Saeiti, entesar.el-saeiti@uob.edu.ly

Abstract

This article would concentrate on hierarchical generalized linear models, including generalized linear mixed-models, which are the extension of linear models. In generalized linear models, the dependent variable assumes every distribution from exponential family distributions, e.g., normal, poisson, binomial, gamma, etc.

The poisson-gamma method was applied, where the dependent variable represents the poisson distribution and the standard error is defined by the gamma distribution. In generalized linear models, several estimation methods have been used. Throughout this study, the hierarchical likelihood estimation method was used to determine the effectiveness of this methodology for both data balanced and unbalanced.

This article compares the Adequacy of poisson-gamma *H*-Likelihood estimation method of mixed effects clustered data models with equal and unequal cluster sizes. This was evaluated in terms of probability of type-I error rate, power and standard error by applying computer simulation. Simulation is performed using different cluster numbers and different cluster sizes. The results show that the performance of the hierarchical likelihood estimation technique provided close approximations in the event of balanced and unbalanced data, while the output of the technique was approximately equivalent in both instances, regardless of cluster size inequality.

Keywords: Hierarchical Generalized Linear Model (HGLM), poisson-gamma *H*-Likelihood, Counting Response, Balanced Clustered, Unbalanced Cluster.

Introduction

Linear models define a continuous response variable as a function of one or more predictor variables. They may help you understand and predict the behavior of complex systems or analyze experimental, financial and biological data.

Linear regression is a statistical method used to construct a linear model. The model describes the relationship between the dependent variable Y (also known as the response), as a function of one or more independent X variables (called predictors).

$$Y = X\beta + \varepsilon, \quad \ldots(1)$$

where β represents linear parameter estimates to be evaluated and ε represents the error terms.

The Generalized Linear Model (GLM) is an extension of the linear model to response variable that follow any probability distribution include the exponential group of distributions.

The exponential family includes useful distributions, for example, normal, binomial, poisson, polynomial, gamma, and others (Leee and Nelder, 2006).

Hypothesis tests applied to the generalized linear model do not require normality of the response variable nor do they require homogeneity of variances. Hence, generalized linear models can be used when response variables follow distributions other than the normal distribution and when variances are not constant. For example, counting data would be appropriately analyzed as a poisson random variable within the context of the generalized linear model.

The Generalized linear mixed model (GLMM) is name as hierarchical generalized linear model. GLMMs can be thought of as an extension of generalized linear models (Lee and Nelder, 2006),

The general form of the model in matrix notation is giving by.

$$Y = X\beta + Zu + \varepsilon \quad (2)$$

McCulloch and Searle (2001) wrote, when studying phenomena within a given period of time or area, the data of any phenomenon will follow the poisson distribution and it is in exponential family. In our paper, it is assumed that the data follow the poisson distribution and the error unit follows the gamma distribution.

From Lee and Nelder's (1996) description of hierarchical models, every distribution in the exponential family has the corresponding distribution, e.g. poisson offset by gamma distribution, binomial distribution offset by beta distribution, normal distribution offset by normal distribution. For more information on hierarchical data structure see El-saeiti (2013, 2014), Lalonde (2009). Cluster data models are frequently used in the field of agricultural, genetic, industrial, medical, biological and even social science experiments. Clustered data or nested data design is an experimental design technique in which data has an implicit hierarchy. The clusters may be balanced or unbalanced, i.e., the number of observations in a cluster (the size of the cluster) is equal or unequal. The unbalanced clustered data may bring up the problem of heterogeneous models which require different variance components, as had been addressed in previous studies for continuous response (El-Saeiti, 2015). In the case of unbalanced clustered data with continuous outcomes in the linear model, El-Saeiti (2015) found that, there was a different dispersions for different clusters sizes. Ac-counting for the different dispersions led to the minimization of mean square error, which was shown through two examples. In this study, the researcher focused on the counting outcomes. When using mixed effects for clustered data with counting outcomes, a preferred model is Hierarchical Generalized Linear Model (HGLM). Lee and Ryan (2017) are concerned with a class of generalized linear mixed models for clustered data, where random effects are mapped solely to cluster structure and are independent between groups; they derive the necessary and sufficient conditions that allow the marginal likelihood of such a class of models to be expressed in closed form. Illustrations are provided using normal, poisson, binomial and gamma distributions; these models are unified under a single umbrella of generalized conjugate linear mixed models, where "conjugate" refers to the fact that marginal likelihood can be expressed in closed form, rather than implying inference through the Bayesian paradigm. Using an explicit marginal likelihood means that these models are more computationally efficient, which can be important in large data environments, with the exception of binomial distribution, so that these models are able to achieve conjugation at the same time and thus be able to accommodate both unit and group level covariates.

Theoretical Background

Poisson-gamma HGLM are members of the hierarchical generalized linear model family (Lee and Nelder, 1996), an extension of the generalized linear model family and the generalized linear mixed model group. For training, Poisson-gamma HGLM is used to characterize historical count data as non-life insurance compensation numbers, among others. It should be remembered that the Poisson gamma HGLM considered at one time follows a negative binomial regression model (Gning, 2013).

Modeling Poisson Data

$$Y_i \sim \text{poisson}(\lambda_i)$$

Then; E(Y_i) = λ_i and Var (Y_i) = λ_i .

The link function must map from (0, ∞) to (∞, ∞). A natural choice is g(μi) = log(μi). For dependent count data (Rönnegård and Shen, 2010) it has been stated that it is common to model a distributed poisson response with a random gamma effect; if no overdispersion is assumed to be conditional on u and thus have a fixed dispersion term; this model may be specified as.

$$E(y_i|\beta, u) \ = \ exp(X_i\beta + Z_i v) \quad ..(3)$$

Lee and Nelder (1996) described the generalized linear model for poisson-gamma hierarchical and the generalized linear mixed form structure of poission. The three pices of
HGLM for poisson- gamma is:

1. $Y_{ij}| u_j \sim \text{poi}\ (\lambda_i, \phi\ \lambda_i))$, $u_i \sim \text{gamm}(\alpha,\gamma_i)$,
2. $\eta = X\beta + Zu$,
3. $\eta = \ln(\lambda_i)$

More details on poisson-gamma model see (Lee and Nelder 1996, 2001).

However, in the clustered count response because the assumption of independence between cluster observations is likely to be violated, a mixed effects clustered counting data model is a useful strategy to account for intracluster correlations in statistical inference, see Hedeker and Gibbons (1994). The purpose of this paper is to compare the performance of the mixed effects clustered data count model with equal and unequal cluster size. Here, the author discusses the probability of type I error rate, the statistical power of the experiment, and the standard error (S.E) by computer simulation study.

Simulation Study

The simplest definition of simulation in science is that it is a numerical method of running trails or tests using computer algorithms instead of conducting a real experiment. Simulation is an approach to modeling random events in such a way that simulated outcomes closely match real-world outcomes, and by studying simulated outcomes, researchers gain knowledge of the real world. In other words, the simulation of a system is the operation of a process model (Maria, 1997). The design can be re-fitted and tested at a lower cost, so simulation will be more realistic.

The function of the prototype can be investigated and thus inferences can be made about the behavior of the real system. Simulation can also be viewed as a method to test the quality of the current or proposed process.

In numerical applications, the word simulation usually involves the random sampling process of the probability distributions. Due to its wide use, this is an important part of the statistical study. This significance occurs in many situations when it is difficult to find statistical diagnosis, or time consuming, or costly to carry out an analysis. Statistical simulation can be used simply by specifying a statistical software that uses random numbers to produce the values of random variables with the desired probability distributions (uniform, normal binomial, etc.) that have been achieved in this research.

For data generation and all simulation steps is included in the Appendix section 'end of this paper'. For more explanation and detail on related simulation studies with different dependent variables and other variables for different purposes, see El-saeiti (2013, 2019). For H-likelihood `poisson gamma HGLM', it was used *hglm* function in *hglm* package for traditional poisson gamma in R throw the simulation steps. Using *hglm* function to get the estimation of parameters β and t-statistic with p value to calculate through simulation.

Results and Discussion

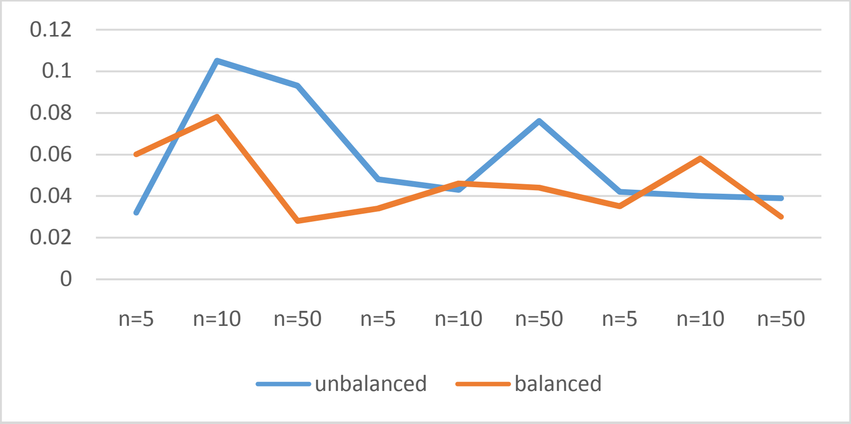
The following tables and diagrams will demonstrate the results obtained from the simulation and display the probability of the type-I error rate in Table (1), the approximation value of the "β " parameters in Table (2), the power in Table (3) and the standard error in Table (4).

Table (1) display the probability of type-I error rate were computed as the proportion of p values less than 0.05 under a null hypothesis $H_0: \beta_1 = 0$ of no treatments effect when we rejected incorrectly.

Table (1): probability of type-I error rate

Cluster	observations	unbalanced	balanced
K=3	n=5	0.032	0.060
	n=10	0.105	0.078
	n=50	0.093	0.028
K=10	n=5	0.048	0.034
	n=10	0.043	0.046
	n=50	0.076	0.044
K=50	n=5	0.042	0.035
	n=10	0.040	0.058
	n=50	0.039	0.030

Fig. (1): Type-I error for poisson gamma



The probability of type-I error rate was acceptable because it was slightly high in some points; generally it was not far away 0.05.

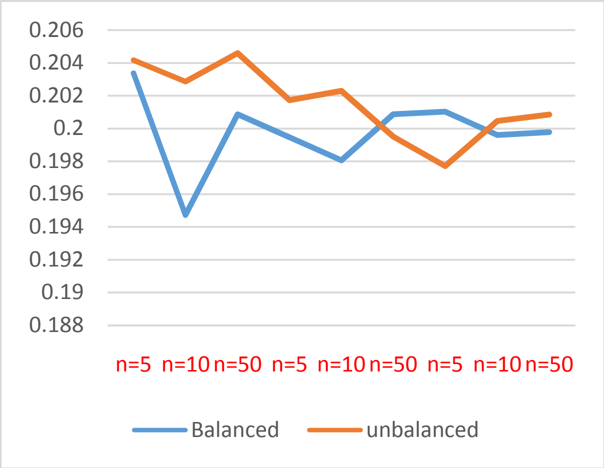
Next table is Table(2); for the simulated sample of size (5,10,50) observations and 3, 10, and 50 clusters; where the actual value is equal to 0.2 for the parameter β_1 , and the value for the β_2 parameter is equal to zero " because there is no X_2 value, it is used only to calculate the power and the probability of type-I error rate"

Table (2): the estimate parameters

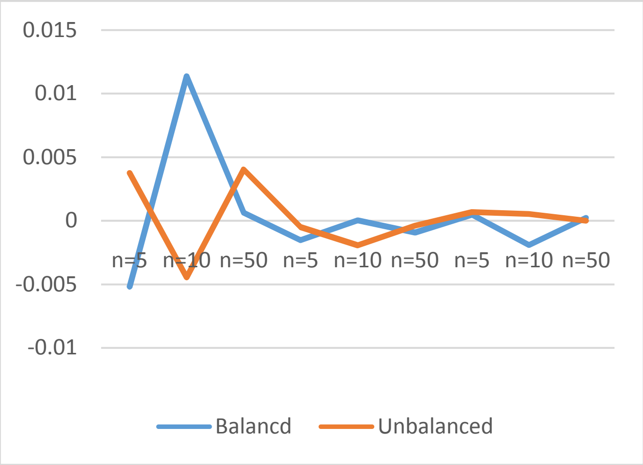
Cluster	Observations	Unbalanced		Balanced	
		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
K=3	n=5	0.2033534	0.003756905	0.2041509	-0.00518115
	n=10	0.1947158	-0.00445098	0.2028573	0.01137586
	n=50	0.2008493	0.004040149	0.2045826	0.00065012
K=10	n=5	0.1994582	-0.00049512	0.2017183	-0.00151689
	n=10	0.1980559	-0.00192974	0.2022803	3.240238e-05
	n=50	0.2008564	-0.00039103	0.1994892	-0.00093368
K=50	n=5	0.2010042	0.0006905077	0.1977052	0.0004996838
	n=10	0.1995827	0.00053609	0.2004524	-0.001911985
	n=50	0.1997564	1.147578e-05	0.2008468	0.0002390147

Table (2) shows that the H-Likelihood estimate was a good estimation method for both cases, since the average of 1,000 replications provided estimates that were very close to the actual values for the parameters. Figs 2.1 and 2.2 included a summary of the predicted values that were close to the actual values.

Fig(2.1): estimate values ($\hat{\beta}_1$)



Fig(2.2): estimate values ($\hat{\beta}_2$)



Next Table (3) demonstrate the power simulated sample of size 5,10, and 50 observations and

The number of : 3, 10, and 50 clusters. Statistical power was computed when rejected hypothesis $H_0 : \beta_2 = 0,$ correctly. Calculate through simulation for 1000 times how many times the test is significant. The power is the proportion of number of rejected correctly.

Table (3) Statistic power

Cluster	observations	unbalanced	balanced
K=3	n=5	0.801	0.808
	n=10	0.963	1.000
	n=50	1.000	1.000
K=10	n=5	1.000	1.000
	n=10	1.000	1.000
	n=50	1.000	1.000
K=50	n=5	1.000	1.000
	n=10	1.000	1.000
	n=50	1.000	1.000

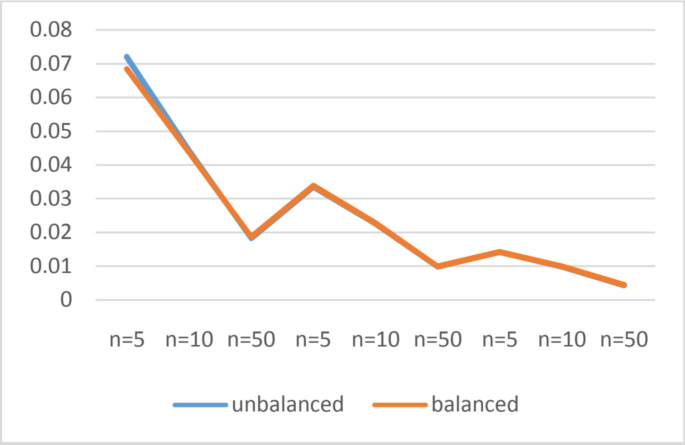
From Table (3) it has been shown that the power values of "probability to accept a null hypothesis that is right" are approximately close to one. The higher power the better method, from the above table hard to decide since the power approximately is 1, and is high for both cases; since the sample size is large for each combination. It is reasonable high power for large sample size, there is no different between both cases in power, both work good according to power for large sample size.

Table (4): stander error (SE) for original simulated sample of size 5,10, and 50 observations and 3,10, and 50 clusters. The stander error was computed as the average of 1000 SEs of the estimates of β_1. The smaller SE represents smaller variability, or greater precision, of the parameter estimates (Heo and Leon, 2005).

Table (4): Stander error for both cases counting data

Cluster	observations	unbalanced	balanced
K=3	n=5	0.07196653	0.06844161
	n=10	0.04438624	0.04374402
	n=50	0.01834041	0.01862745
K=10	n=5	0.03354806	0.03383271
	n=10	0.02272336	0.02271749
	n=50	0.00991545	0.00993959
K=50	n=5	0.01423205	0.01417949
	n=10	0.00995692	0.00992440
	n=50	0.00443169	0.00443549

Fig.(4) Standard error for balanced and unbalanced data.



From Table (4) and graph (4) above, it can be seen that there is no difference in the standard error for HGLM poisson game in balanced and unbalanced counting data

Discussion

In this article, we looked at the generalized linear mixed-models, which are the extension of linear models. It is understood that many other studies have studied a problem with unbalanced data or incomplete information, which may lead to a heterogenetic problem. The heterogenetic issue was not discussed here by the use of the hierarchical probability estimation model.

The process has impartial and very similar outcomes in two situations that are balanced and unbalanced. The lack of meaning and the imbalanced model will therefore not be a concern by using the poisson-gamma *H*-Likelihood estimation.

The hierarchical probability estimation approach has been concluded to be able to solve heterogenetic problems in future studies.

As stated earlier, this study's main objective was the efficiency of the *H*-Likelihood estimation approach for unbalanced cluster data models. *H*-Likelihood estimation approach the system for unbalanced clustered count data models is recommended in order to avoid heterogeneity problems.

References

El-Saeiti, I. N. (2014) "*Performance of Mixed Effects for Clustered Binary Data Models". *AIP Conference Proceedings 1643, 80

EL-Saeiti, I. N. (2015). "Messy data in heteroscedastic models case study: Mixed nested design". LAP Lambert Academic Publishing.

EL-Saeiti, I. N. (2019). An adjusted scale binomial Beta H-Likelihood estimation method for unbalanced clus-tered binary response models. *Libyan Journal of Science & Technology*; Vol. (10:1) 20-22.

EL-Saeiti, I. N.(2013): *"Adjusted variance components for unbalanced clustered binary data models".* Ph. Doctoral "University of Northern Colorado."

Gning, L.(2013). On the existence of maximum likelihood estimators in Poisson-gamma HGLM and negative binomial regression model. *Electronic Journal of Statistics;* Vol. (7), 2577–2594

Heo, M. and Leon, A. (2005). Performance of a mixed effects logistic regression model for binary outcomes with unequal cluster size. *Biopharmaceutical Statistics*,15:513-526.

L.Gning and D. Pierre-Loti-Viaud.(2012): On the existence of maximum likelihood estimators in poisson-gamma HGLM and negative binomial regression model.

Lalonde, T. L. (2009). Components of overdispersion in hierarchical generalized linear models. Dissertations " University of Northern Colorado".

Lee, Y., & Nelder, J. A. (2006). Double hierarchical generalized linear models. Journal of the Royal Statistical Society, Series B (Methodological), 55, 139-185.

Lee,Y. and Nelder, J.(1996): Hierarchical Generalized Linear Models Journal of the Royal Statistical Society, Series B (Methodological), 58 (4), 619-678.

Maria, A.(1997): "Introduction to Modeling and Simulation", Proceedings of the 1997 Winter Simulation Conference.

McCulloch, C.E., and Shayle, R.S.(2001): *Generalized, linear, and mixed* models. NY: John Wiley & Sons, Inc.

Rönnegård, L., Alam,M. and Shen,X. (2010) hglm Package (Version 2.0) Package Maintainer

Appendix

```
mydata=function(seed){
set.seed(seed)
   beta0   = 1
   beta1   = 0.2
   beta2   = 3.1
########## for poi-gam###
 n.clus <- 2         #No. of clusters
 n.per.clus <- 5     #No. of obs. per cluster for equal
 sigma2_u <- 0.2     #Variance of random effect
 sigma2_e <- 1       #Residual variance
 sigma1<- 2
nn <- n.clus*n.per.clus
beta=matrix(c(beta0,beta1,beta2),3,1)

y=matrix(0,nn,1)
X=matrix(c(rep(1,nn),rep(0,nn),rep(0,nn)),nn,3)
Z=matrix(0,nn ,n.clus)
a <- rnorm(n.clus, 0, sqrt(sigma2_u))
e <- rnorm(nn, 0, sqrt(sigma2_e))

 ## GENERATE X-VALUES FROM NORMAL DISTIRBUATION##

        X[,2] =rnorm(nn,3,sigma1)
        X[,3]=rpois(nn,3)

X_d <- matrix(c(rep(1,nn),rep(0,nn),rep(0,nn)),nn,3)
Z <- diag(n.clus)%x%rep(1, n.per.clus)
u <- rgamma(n.clus,1)
eta <- exp(beta0+beta1*X[,2]+Z%*%u)
y <- rpois(length(eta), eta)

    list( X=X, y=y,u=u,Z=Z, X_d=X_d)
}
############################################
#       POWER FOR H-LIKELIHOOD FUNCTION  #
#                By using hglm function          #
########================================#
library(MASS)
library(hglm)
simA= function (N1){
  set.seed(1234)
alpha       <-  0.05
b21count    <-  0
b22count    <-  0
S.E2        <-  matrix(0,nrow=N1, ncol=1)
b.E21        <-  matrix(0,nrow=N1, ncol=1)
b.E22        <-  matrix(0,nrow=N1, ncol=1)
seeds=rnorm(N1,0,50)
set.seed(seeds)
for(i  in 1:N1)
{
datta= mydata(seeds[i])
X=datta$X
y=datta$y
X_d=datta$X_d
Z=datta$Z
 #========================h-likelihood method ==========================#
R <-  gamma.pois <- hglm(y = y, X = X, Z = Z,
 X.disp = X_d, family = poisson(link = log), rand.family =
 Gamma(link = log))
SS= summary(R)
betas <- R$fixef
se    <- R$SeFe
zval  <- betas / se
pval   <- 2 * pnorm(abs(zval), lower.tail = FALSE)
S.E2[i,] <- se[2]
b.E21[i,] <- betas[2]
b.E22[i,] <- betas[3]
       p21 = pval[2]
    if(p21 < alpha){b21count = b21count+1}

       p22 = pval[3]
    if (p22 < alpha){b22count = b22count+1}
}
typeI2=b22count/N1
power2=b21count/N1
se2 <- sum(S.E2)/N1
be21 <- sum(b.E21)/N1
be22 <- sum(b.E22)/N1

list(SS=SS, be21=be21,be22=be22,power2=power2,typeI2=typeI2,se2=se2)
}
```