PERFORMANCE OF MIXED EFFECTS FOR CLUSTERED COUNTING DATA MODELS

Intesar N. El-Saeiti[1*] and Khalil Mostafa ALsawi[2]

[1,2] Department of Statistics, Faculty of Science, University of Benghazi, Libya.
[*] entesar.el-saeiti@uob.edu.ly

Abstract

This article is concerned with hierarchical generalized linear models. It includes generalized linear models and generalized linear mixed models, which are related to linear models.

In generalized linear mixed models, the dependent variable and the standard error follow any distribution from the exponential family, e.g. normal, Poisson, binomial, gamma, etc. We studied counting data, and then use the Poisson-gamma model,where the dependentvariable follows the Poisson distribution and the standard error follow the gamma distribution. Several estimation techniques can be used for generalized linear mixed model. In this paperthe hierarchical likelihood estimation technique was used to prove the performance of H-likelihood methodwhen thecounting data were balanced or unbalanced.

Real data were used to test the performance of Poisson-gamma H-likelihood estimation method in case of balanced and unbalanced counting data.When real data used in the past research for another problem, it was noticed that the performance of the hierarchical likelihood estimation technique gave a close approximations in the event of balanced and unbalanced counting data, and the output of the technique was approximately equivalent in both instances.

Keywords: Mixed Effects; Poisson-Gamma H-Likelihood Model; Clustered Counting Data; Balanced and Unbalanced Clustered.

Introduction

Linear models describe a continuous response variable as a function of one or more predictor variables. It is used to understand and predict the behavior of complex systems, and to analyze experimental, financial, and biological data.

Linear regression is a statistical method used to create a linear model that describes the relationship between a dependent variable Y (also called the response) as a function of one or more independent variables X (called the predictors). The general equation for the linear model is:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (1)$$

Where β represents the linear parameter estimate to be computed and ε represents the error terms. There are several types of linear regression (Simple; Multiple; Multivariate linear regression).

The extension of linear models is a hierarchical generalized linear model (HGLM) in which theresponse variables follow any probability distribution includedin the exponential family. The exponential family includes useful distributions such as normal, binomial, Poisson, polynomial, gamma. No hypothesis tests are applied to the generalized linear model normal response variable and it does not require homogeneity of differences; for more details see Lalonde (2009) & El-saeiti (2013).

The Generalized linear mixed model (GLMM) is also called the hierarchical generalized linear model (HGLM). (HGLM) 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:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{ZU} + \boldsymbol{\varepsilon} \qquad \textbf{(2)}$$

McCulloch and Searle (2001) mentioned that 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. Lee et.al (2006) name the hierarchical models for each distribution in the exponential family as a corresponding distribution, for example Poisson offset by gamma distribution, binomial distribution offset by beta distribution, normal distribution is offset by normal distribution.

Clustered data models, which research structures with hierarchical (nested) classifications, are commonly used in agricultural, genetic, manufacturing, clinical, biological and even social science field experiments. Unbalanced clustered data raise the issue of heterogeneous models that require different variance components (EL-Saeiti,2015). It was found that there were different dispersions of different cluster sizes in the case of unbalanced clustered data with continuous effects in the linear model. Moreover, it was noticed that counting the various dispersions minimized the mean square error, which was shown in two cases. Counting the results is the main concern here. When using mixed effects of clustered data with counting the results, the Generalized Hierarchical Linear Method (HGLM) is the appropriate model. Here the data are counting ones, which follow a Poisson distribution and the error unit will follow the gamma distribution.Missing some of the data can be there;these missing data can cause serious problems.Rubin (1976) mentioned that missing data (or missing values) are defined as the data values that are not stored for an interest observation variable. The problem of missing values can be considered one of the most serious problems in many databases with different types of recorded variables and different data sources. The problem can occur due to lack of information forcertain reasons, or no responses are provided for the questionnaires. The existence of missing data may impact on the results of any statistical study based on such information as the missing values would alter the total pattern (image) of what the data reflects

at both levels; the exploratory and inferential statistics. Missingdata decreases the statistical power, which relates to the probability of rejectingor accepting the null hypothesis of the experiment. Missing data can cause bias in the estimation of parameters, and it may complicate the analysis of the study. Each of these inconsistencies may constitute a threat to the credibility of the trials and may lead to incorrect conclusions. On the other hand, lack of information can increase dispersion or contrast between layers, which can lead to heterogeneity problems. So, since in GLMM do requireneithernormality nor heterogeneity, the researchers can use it to know H- Likelihood estimation method with balanced data and unbalanced data.

Data Collection and Analysis

Data were initially collected from the US National Medical Expenditure Survey (NMES) for 1987/88, which are available in the Journal of Applied Econometrics information collection.

Deb and Trivedi (1997) clarified the actual data gathered from patients for health care purposes. These data are for 4406 individuals aged 66 and over who are covered by Medicare, which is a public insurance system. Data were designed for R package supporting Kleiber and Zeileis (2008) and is also available as DebTrivedi.rda in the Journal of Statistical Software together with Zeileis (2006).

Here, we adopted the number of medical office visits as the dependent variable and used the health status variables hosp (number of hospital stays), health (self-perceived health status), numchron (number of chronic conditions), as well as the socio-economic variables age, school (number of years of education) and private (private insurance) as predictors.

The R software was used to figure out how the (hglm) approach is effective to calculate the counting of the balanced data and unbalanced data. We followed the steps mentioned in Zeileis's paper. In addition, we used the method of (hglm) to estimate the data in full. Then we used the library (miss Forest) in the program R to delete the random data by prod NA (delete the following; 5%, 10%, 15%, 20%); this deletion was in the dependent variable (ofp). After the deletion process, the program placed the symbol (Na)instead of each deleted value since it is not possible to apply (hglm). Therefore, we used the command (na.omit) to delete all the values containing (Na). The data is unbalanced and applied (hglm).

Results and Discussion

The following table shows theresults of a balanced and unbalanced design with missing percentage in the variables (5%, 10%, 15% and 20%). The comparison is done by looking at unbiased parameters and standard error for fixed and random parts.

Table (1): Result of the Fixed Effects Estimates

Summary of the Fixed Effects Estimates					
	Balanced	**Unbalanced**			
	Expected Value	**Expected Value Missing 5%**	**Expected Value Missing 10%**	**Expected Value Missing 15%**	**Expected Value Missing 20%**
(Intercept)	1.353867	1.306338	1.456327	1.304676	1.266936
Numchron	0.147007	0.146573	0.142249	0.148410	0.140636
Gendermale	-0.097134	-0.099085	-0.092532	-0.076437	-0.059536
Blackyes	-0.046460	-0.051993	-0.074086	-0.076147	-0.082763
Hosp	0.159719	0.165138	0.157892	0.162438	0.165676
Age	-0.046535	-0.042469	-0.056122	-0.045328	-0.040360
Marriedyes	-0.055903	-0.055464	-0.066275	-0.073873	-0.090968
School	0.022631	0.021844	0.022112	0.023404	0.024780
Employedyes	0.037279	0.035841	0.043149	0.064870	0.063906
privinsyes	0.208899	0.228606	0.186592	0.216578	0.218232
regionnoreas	0.135856	0.134932	0.141298	0.138793	0.164833
regionother	0.023227	0.021085	0.026506	0.025196	0.023088
regionwest	0.159638	0.164768	0.158276	0.184506	0.192220
emer	0.016421	0.022261	0.023252	0.027402	0.034175

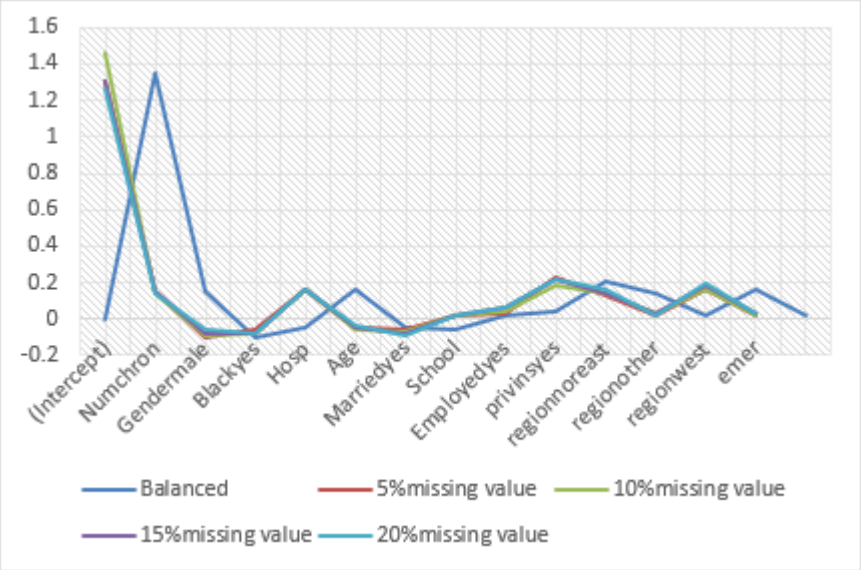


Figure (1): Expected Value of the Fixed Effects Estimates.

From Table (1) it is clear that the results were very close whether the counting data were balancedor unbalanced; Figure (1) for convenient views that values are approximately same.

Table (2): Result of the Standard Error for Fixed Effect Variable

	Balanced	**Unbalanced**			
	Std. Error	**Std. Error Missing 5%**	**Std. Error Missing 10%**	**Std. Error Missing 15%**	**Std. Error Missing 20%**
(Intercept)	0.265284	0.266565	0.275265	0.267425	0.276911
Numchron	0.010578	0.010809	0.011175	0.011467	0.011878
Gendermale	0.032509	0.033330	0.034360	0.035581	0.036626
Blackyes	0.051217	0.052610	0.054552	0.056761	0.058930
Hosp	0.015915	0.016037	0.017055	0.017523	0.018019
Age	0.024157	0.024774	0.025623	0.026174	0.026953
Marriedyes	0.033083	0.033939	0.034981	0.036078	0.037237
School	0.004363	0.004466	0.004619	0.004754	0.004903
Employedyes	0.049943	0.051577	0.052346	0.054086	0.055917
privinsyes	0.040370	0.041765	0.042396	0.044305	0.045713
regionnoreas	0.043341	0.044372	0.045692	0.047229	0.048551
regionother	0.038367	0.039381	0.040383	0.041914	0.043487
regionwest	0.043694	0.044741	0.045890	0.047216	0.048749
emer	0.017304	0.017420	0.018757	0.018163	0.018661

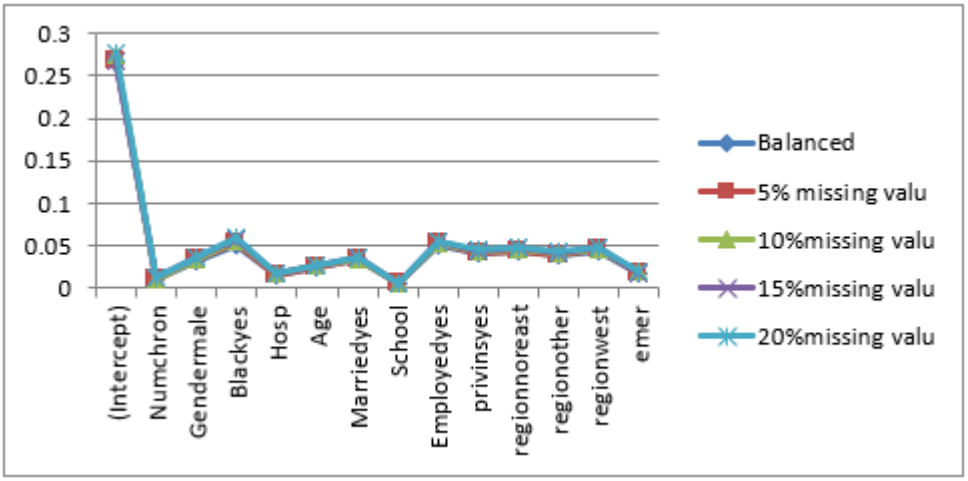


Figure (2): Standard error of the fixed effects estimates.

Table (2) and explanation in Figure (2) show standard error result values, which are very close to each other.

Next tables and graphs explain the outcomes values and figures for predictors of random variable "Health poor; Health average; health excellent" respectively.

Table (3): outcome of the predictor values for random effect variables

	Balanced	**Unbalanced Predictor Values When:**			
	predictor values	**Missing 5%**	**Missing 10%**	**Missing 15%**	**Missing 20%**
Healthpoor	1.2921	1.2784	1.2917	1.2529	1.2721
Healthaverage	1.0068	1.0163	1.0127	1.0205	1.0091
healthexcellent	0.7011	0.7052	0.6957	0.7267	0.7188

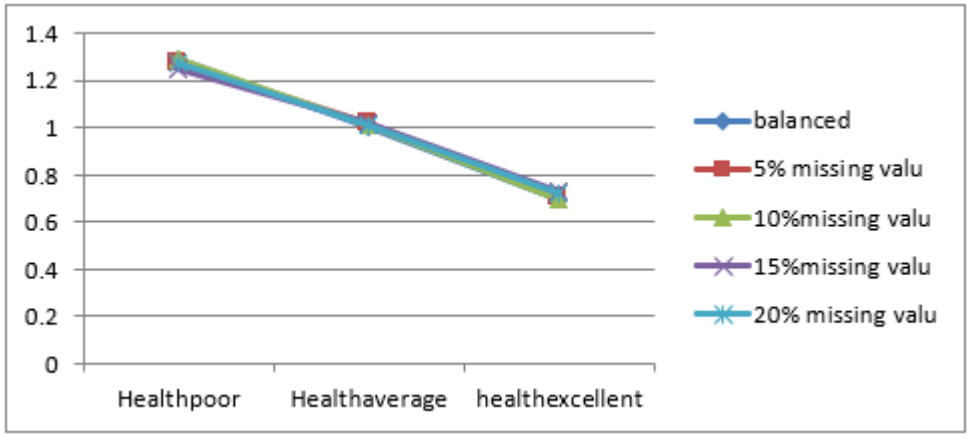


Figure (3): Predictor Values of the Random Effects Estimates.

Table (4): Outcome of the Stdandrad Error Values for Random Effect Variable

	Balanced	**Unbalanced Std. Error When:**			
	Std. Error	**Missing 5%**	**Missing 10%**	**Missing 15%**	**Missing 20%**
Healthpoor	0.1792	0.1752	0.1818	0.1622	0.1693
Healthaverage	0.1787	0.1746	0.1812	0.1614	0.1686
healthexcellent	0.1853	0.1817	0.1885	0.1693	0.1768

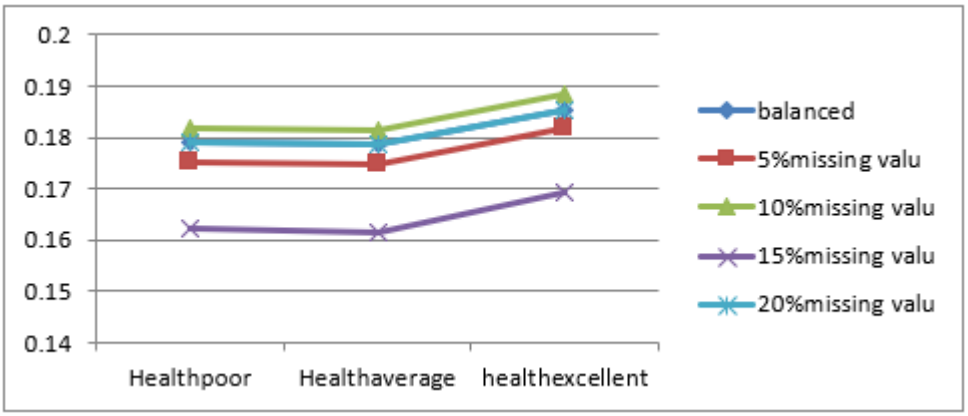


Figure (4): Standard Error of the Random Effects Estimates.

It is very clear from Table (3) and Table (4) that the effects of the values were identical and when missing 15% the result was slightly different in Table (4), but it is not a matter of concern. The value of points in Figures (3) and (4) are very similar to each other.

Conclusion

Based on the results obtained when the H-likelihood estimation method was tested, the Poisson-gamma model, which was used as the dependent variable "accounting variable", followed the distribution of Poisson, and the standard error followed the distribution of gamma. GLMM has several estimation methods; the H-likelihood estimation method was used to estimate its effectiveness with both balanced and unbalanced counting data. It is known that many other studies dealt with unbalanced data problems such as unbalanced or missing data that may led to heterogenetic. In this work, the researchers used the hierarchical likelihood estimation method, which proved that the heterogenetic problem did not surface. Moreover, the method had unbiased and very close results in both cases balanced and unbalanced data. Therefore, through using this Poisson-gamma H-likelihood estimation the missing value and unbalanced model will not be a problem.

It was concluded the H-likelihood estimation method has adequacy and may solve heterogenetic problem in future studies. Then, it is worthwhile applying H-likelihood method in future studies in case of the counting data are unbalanced.

References

- Deb, P. and Trivedi, P.K. (1997). Demand for Medical Care by the Elderly: A Finite Mixture Approach. Journal of Applied Econometrics, 12, 313-336.

- El-Saeiti Intesar (2013). “Adjusted variance components for unbalanced clustered binary data models”. Ph. Doctoral. Dissertations.

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

- Kleiber C, Zeileis A (2008). Applied Econometrics with R. Springer-Verlag, New York.

- Lalonde, T. L. (2009). Components of overdispersion in hierarchical generalized linear models. Ph. Doctoral. Dissertations.

- Lee, Y. Nelder., J. and Pawitan, Y. (2006). Generalized linear models with random effects. Boca Raton: Chapman & Hall.

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

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

- Rubin, D. B. (1976). Inference and missing data. Biometrika, 63, 581–592.

- Zeileis A (2006). Object-oriented Computation of Sandwich Estimators. Journal of Statistical Software, 16(9), 1{16. URL http://www.jstatsoft.org/v16/i09/.