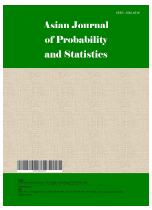

Asian Journal of Probability and Statistics

8(3): 31-37, 2020; Article no.AJPAS.59281
ISSN: 2582-0230

Comparison of Generalized Linear Model and Generalized Linear Mixed Model – An Application to Low Birth Weight Data

Michael Fosu Ofori[1*], Stephen B. Twum[2] and Jackson A. Y. Osborne[3]

[1]Department of Mathematics & Statistics, Kumasi Technical University, Ghana.
[2]Department of Mathematics, University for Development Studies, Navrongo, Ghana.
[3]Faculty of Applied Science, Methodist University College, Accra, Ghana.

Authors' contributions

This work was carried out in collaboration among all authors. Author MFO designed the study, performed the statistical analysis, wrote the protocol and wrote the first draft of the manuscript. Authors SBT and JAYO managed the analyses of the study. Author MFO managed the literature searches. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJPAS/2020/v8i330208
Editor(s):
(1) Dr. Jiteng Jia, Xidian University, China.
Reviewers:
(1) Abhiram Dash, Odisha University of Agriculture & Technology, India.
(2) Ajenikoko Ganiyu, Ladoke Akintola University of Technology, Nigeria.
Complete Peer review History: http://www.sdiarticle4.com/review-history/59281

Original Research Article

Received: 19 May 2020
Accepted: 24 July 2020
Published: 22 August 2020

Abstract

Background: Generalized Linear models are mostly fitted to data that are not correlated. However, very often data that are collected from health and epidemiological studies are correlated either as a result of the sampling methods or the randomness associated with the collection of such data. Therefore, fitting generalized linear models to such data that produce only fixed effects could lead to over dispersion in the model estimates.
Objectives: The objective of this study is to fit both generalized linear and generalized linear mixed models to a correlated data and compare the results of the two models.
Methods: Logistic regression is employed in fitting the generalized linear model since the dependent variable in the study is bivariate whilst the GLIMMIX model in SAS is used to fit the generalized linear mixed model.
Results: The generalized linear model produces over dispersion with higher errors among the parameter estimates than the generalized linear mixed model.
Conclusion: In dealing with a more correlated data, generalized linear mixed model, which can handle both fixed and random effects, is preferable to generalized linear model.

**Corresponding author: E-mail: mikeoffos@yahoo.com, mikeoffos70@gmail.com;*

Keywords: Generalized linear mixed model; logistic; correlated; fixed effects; random effects.

1 Introduction

The GLIMMIX procedure is an add-on procedure in SAS/STAT software which is a product in SAS 9.1 on the Windows platform. This procedure is currently downloadable for the SAS 9.1 release from Software Downloads at support.sas.com.

The GLIMMIX procedure performs estimation and statistical inference for generalized linear mixed models (GLMMs). A generalized linear mixed model is a statistical model that extends the class of generalized linear models (GLMs) by incorporating normally distributed random effects. A GLM can be defined in terms of several model components:

- a linear predictor η that is a linear combination of regression coefficients: $\eta_i = X_i'\beta$
- a link function $g(.)$ that relates the mean of the data to the linear predictor, $g(E[Y_i]) = \eta_i$
- a response distribution for Y_i from the exponential family of distributions

The exponential family of distributions is very broad and contains many important distributions. For example, the binary, binomial, Poisson, negative binomial, normal, beta, gamma, and inverse Gaussian distribution are members of this family. A special case of the generalized linear model arises when the Yi are normally distributed and the link function is the identity function. The resulting models are linear regression and analysis of variance models with normal errors.

Generalized linear models apply when the data are uncorrelated. However, in many studies, like in our present study, observations exhibit some form of dependency. For example, measurements of different attributes are taken from the same mother, observations are collected over time, sampling or randomization is carried out hierarchically, and so forth.

2 Objectives

The main objective is to use the GLIMMIX procedure [1,2] to propose a joint modeling of correlated binary and continuous data as pertains in our present study where we have both binary variables such as antenatal care and locality, and continuous variable like age in the data.

Other specific objectives are;

1. To propose a joint model involving correlated binary and continuous data using the GLIMMIX procedure
2. To compare the GLM with the model by the GLIMMIX procedure using the same data.

3 Methods

The study employed data based on the third round Multi Indicator Cluster Survey (MICS) conducted by Ghana Statistical Service [3]. The issues regarding this data include the following;

1. The data exhibit some randomness (e.g., selection of the households, women were asked several questions spanning between two years – survey period, the classifications for ages and parity were also done randomly)
2. There is high correlation between variables such as age and parity (number of children ever born)
3. The data contain both categorical and continuous variables (e.g., antenatal care and mothers' age)
4. The data show a prediction for a binary variable.

To address the deficiencies of generalized linear models and linear mixed models outlined above, we employ the GLIMMIX procedure to model our data so as to come up with a more efficient model compared with the generalized linear model. A model that can address both fixed and random effects issues since the data is non normal and exhibit correlation among some parameters especially age and parity. PROC MIXED is another way of dealing with both fixed and random effects; however, it is appropriate only when the data is normal unlike our present study where the data being dealt with is not normal in which the prediction for the binary variable is bounded.

Our earlier investigation dealt with a binary outcome modelling approach using PROC GENMOD with the link function, [4,5]. The analysis is now extended to PROC GLIMMIX which involves both marginal (R-side) and random (G-side) effects – random intercept model. The GLIMMIX technique fits statistical models to data with non-constant variability or correlations and where the response is not necessarily normally distributed by using RMPL. The problem is that GLMs are fitted to uncorrelated data that have only fixed effects.

The questions are; What does one do when there is correlation and random effects in the data? Which model do we apply to the data? Should we still fit a GLM to the data?

The GLMs are applied on uncorrelated data. (It assumes a fixed linear process fitted to normal or non-normal data but has no random effect). However, in many studies, as is the case in our present study, there may be some amount of dependency among observations. For instance, measuring different attributes from the same mothers over a period of time may cause dependency. In fact, in our earlier analysis, age and parity are found to be highly correlated.

The GENMOD procedure in SAS fits GLMs for independent data using maximum likelihood. The procedure also handle data that are correlated using the marginal GEE approach of [6-10].

The LMMs are improvement on GLM which assumes a linear process with fixed and random components fitted to normal data. However, the LMMs is handicapped when the data is non-normal (especially when the response variable is binomial as in our present study and in many health and epidemiology studies).

The models fit by the GLIMMIX technique extend the GLM by allowing correlations among the responses. This can be achieved with an inclusion of random effects in the linear predictor and/or by directly modeling the correlations among the data.

The GLIMMIX technique differentiate the two procedures as “G-side” and “R-side” random effects. This term draws on a common specification of the linear mixed model [11-19].

$$Y = X\beta + Z\gamma + e \tag{1}$$

The G-side random effects are computed the same way by adding random effects to the linear predictor. This provides a model of the form

$$(E[Y|\gamma]) = X'\beta + Z'\gamma \tag{2}$$

A model with only R-side random effects, [20-23] is known as a marginal model in that no random effects exists on which the response could be conditioned. In such a model, the mean is specified as

$$(E[Y]) = (\mu) = x'\beta \tag{3}$$

When these elements are combined, we can represent the models fit by the GLIMMIX approach as follows:

$$[Y|\gamma] = g^{-1}(X\beta + Z\gamma) = g^{-1}(\eta) = \mu$$

$$var[\gamma] = G$$

$$[Y|\gamma] = D^{1/2}RD^{1/2} \quad (4)$$

where $g^{-1}(.)$ represents the inverse link function [24,25].

3.1 Justification for random intercept model

1. The linear mixed model assumes that the random effects follow a normal distribution and the variance is not a function of the mean. The constant variance assumption is not applicable when analysing a zero/one trait, such as LBW (0) or NWB (1). Here, the response variable is Binomial. Hence, for a predicted LBW incidence, the variance is $(1-\mu)$, which is a function of the mean.
2. The normality assumption does not hold for a binary trait. The result is a random variable that can take two values only, one or zero. Contrary, the normal distribution is a bell-shaped curve that can take any real number.
3. Predictions from linear mixed models can handle any value whilst predictions for a binary variable is bounded (0,1) or cannot take negative values for a count variable.
4. More importantly, if data are correlated, a standard GLM will show over-dispersion in relation to the binomial distribution. The best way to deal with this over-dispersion is by adding the G-side random effects to indirectly model the correlation which is influenced by distributing the random effects.

As a result of these shortcomings associated with the GLMs and the LMMs, coupled with the dependency (correlation) among the variables, for instance age and parity, there is the need to come up with a model that will address the identified challenges.

4 Results

Table 1 compares the Generalized linear model and the Generalized linear mixed model (or Random intercept model). Since the data collection processes involved some randomness we sought to explore the data by comparing the GLM which deal with uncorrelated data and GLMM which invokes correlation processes. It is realize in both cases, the random intercept model using the GLIMMIX procedure predicts better than the GLM as shown in Table 1. This means that in a well correlated data such as in epidemiology and health studies where data are mostly correlated as is the case in our present study, random intercept model using the GLIMMIX procedure is the preferred choice. Comparing the two models, it can easily be deduced from the fit statistics that the errors are far lower with the proposed model (GLMM) than the GLM. The greater variability exhibited by the parameter estimates in the GLM indicates over dispersion. The GLMM on the other hand shows perfect correlation between age and parity, fitting very well to the data. This shows that fitting a GLM to a more correlated data will lead to over dispersion in the binomial model.

The results from our random intercept model using the GLIMMIX procedure produced better results with minimal errors compared with the logistic regression model. This indicates that the randomness (dependency) among some of the covariates is an issue among epidemiology and health data. This stands to reason that in dealing with such data sets, both fixed effects and random effects must be looked at especially where normality assumptions are violated as is the case in our present study.

5 Discussion

The random intercept model enhances the results of the generalized linear model by taking into consideration the randomness in the data. The results show some dependencies among some variables, especially parity and age. The model indicates greater variability in the GLM than the random intercept model. Comparing both GLM and GLMM, it is realized in both cases, the random intercept model using the

GLIMMIX procedure predicts better than the GLM as shown in Table 1. This means that in a well correlated data such as in epidemiology and health studies where data are mostly correlated as is the case in our present study, random intercept model using the GLIMMIX procedure should be the preferred choice. The GLMM also provides age especially age above 35 as significantly associated with low birth weight. The GLM however, shows over dispersion in the results as there are greater variability among the parameter estimates compared with the random intercept model, which shows perfect correlation between age and parity.

Table 1. Comparison of GLM and GLMM

Parameters	Generalized linear model					Generalized linear mixed model			
	Estimates (β)	Std. error	Hypothesis test		Exp. (β)	Estimates (β)	Std. error	t Value	Pr>|t|
			Wald Chisq	Sig.					
Intercept	-0.912	1.7216	0.281	0.596	0.402	0.3844	0.19750	1.95	0.0519
locality=1	0.066	0.1688	0.153	0.696	1.068	4.16E-08	0.00003	0	0.9990
Locality =2	0	-	-	-	1	0	.	.	.
ANC=1	0.151	1.1049	0.019	0.892	1.162	0.01234	0.12730	0.10	0.9233
ANC=2	0	-	-	-	1	0	.	.	.
Age	0.205	0.1026	4.008	0.045	1.228	0.02830	0.011070	2.56	0.0144
Age squared	-0.003	0.0016	3.196	0.074	0.997	-0.00036	0.000168	-2.14	0.0379
Parity1*age	-0.049	0.0274	3.234	0.072	0.952	-0.00134	0.002053	-0.65	0.5163
Parity2*age	-0.035	0.0271	1.645	0.200	0.966	-0.00134	0.002053	-0.65	0.5163
Parity3*age	-0.028	0.0268	1.118	0.290	0.972	-0.00134	0.002053	-0.65	0.5163
Parity4*age	-0.025	0.0261	0.944	0.331	0.975	-0.00134	0.002053	-0.65	0.5163
Parity5*age	-0.013	0.0264	0.253	0.615	0.987	-0.00134	0.002053	-0.65	0.5163
Parity6*age	-0.014	0.0268	0.272	0.602	0.986	-0.00134	0.002053	-0.65	0.5163
Parity7*age	-0.029	0.0255	1.278	0.258	0.972	-0.00134	0.002053	-0.65	0.5163
Parity8*age	-0.020	0.0346	0.348	0.555	1.021	-0.00134	0.002053	-0.65	0.5163
Parity9*age	-0.036	0.0275	1.740	0.187	0.964	-0.00371	0.002696	-1.38	0.1761
Parity10*age	0				1	0			.
(Scale)	1								
Model fit statistics									
	-2Res Log Likelihood		**AIC**		**AICC**	**BIC**	**CAIC**	**HQIC**	
GLM	236.428		500.857		501.176	573.589	587.589	-	
GLMM	340.54		372.540		372.960	455.50	471.500	403.650	

6 Conclusion

The random intercept model enforced the importance of giving credence to randomness in epidemiology and health data such as what pertains in our present study, which usually have some correlations among some of the covariates since information are collected on individual subjects over a period of time. The results of this model identified the dependency among some variables and compared to the logistic model provided better results with fewer errors. The GLM also showed over dispersion with greater variability in the parameter estimates than the GLMM, giving an indication that GLM performs poorly when fitted to a more correlated data.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] SAS Institute Inc. Select Chapters in *SAS/STAT® 9.3 User's Guide*. Cary, NC: SAS Institute Inc.; 2012.

[2] Diggle PJ, Liang KY, Seger SL. Analysis of longitudinal data. Oxford University Press, New York; 1995.

[3] Ghana Statistical Service. Multiple Indicator Cluster Survey; 2011.

[4] User's Guide The GLIMMIX Procedure SAS Institute Inc. SAS/STAT® 13.1 Cary, NC: SAS Institute Inc.; 2013.

[5] Tyler Smith, Besa Smith. PROC GENMOD with GEE to analyze correlated outcomes data using SAS; 2013.

[6] Zeger SL, Liang KY. Longitudinal data analysis using generalized linear models. Biometrica. 1986;73:13-22.

[7] McCullagh P, Nelder J. Generalized linear models. Chapman and Hall, London, 2nd Edition; 1989.

[8] Edward Gbur, et al. Analysis of generalized linear mixed models in the agricultural and natural resources sciences. Madison WI 53711-5801, USA; 2012.

[9] De Silva NH, Gea L, Lowe R. Genetic analysis of resistance to *Pseudomonas syringae* pv. actinidiae (Psa) in a kiwifruit progeny test: An application of generalised linear mixed models (GLMMs). SpringerPlus. 2014;3:547.

[10] Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White JS. Generalised linear mixed models: A practical guide for ecology and evolution. Trends in Ecol and Evolution. 2009;24(3):127-135.

[11] Gilmour AR, et al. The analysis of binomial data by a generalized linear mixed model. Biometrika. 1985;72:593–599.

[12] Schall R. Estimation in generalized linear models with random effects. Biometrika. 1991;78:719–727.

[13] Wolfinger R, O'Connell M. Generalized linear mixed models: A pseudo-likelihood approach. J. Statist. Comput. Simulation. 1993;48:233–243.

[14] Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. J. Am. Stat. Assoc. 1993;88:9–25.

[15] Raudenbush SW, et al. Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. J. Comput. Graph. Statist. 2000;9:141–157.

[16] Pinheiro JC, Chao EC. Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. J. Comput. Graph. Statist. 2006;15:58–81.

[17] Lindsey JK. Applying generalized linear models. Springer; 1997.

[18] Benjamin MB, Mollie EB, Connie JC, Shane WG, John RP, Stevens MHH, Jada-Simone SW. Generalized linear mixed models: A practical guide for ecology and evolution; 2008.

[19] Donald Hedeker. Generalized linear mixed models; 2005.

[20] Kelvin YKW, Anthony KYC. Robust estimation in generalized linear mixed models; 2002.

[21] Schabenberger O. Introducing the GLIMMIX procedure for generalized linear mixed models. SUGI 30 Proceedings, Philadelphia, Pennsylvania, US. 2005;196-130. Google Scholar

[22] Schabenberger Oliver. Introducing the GLIMMIX procedure for generalized linear mixed models; 2005.

[23] Schabenberger O. Growing up fast: SAS1 9.2 enhancements to the GLIMMIX procedure. SAS Global Forum. 2007;177.
Available:www2.sas.com/proceedings/forum2007/177-2007.pdf

[24] Littell RC, et al. SAS for mixed models. (2nd Edn), SAS Publishing; 2006.

[25] Booth JG, Hobert JP. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. J. R. Stat. Soc. Ser. B Methodological. 1999;61:265–285.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://www.sdiarticle4.com/review-history/59281