Sandwich Variance Estimation for random effect misspecification in Generalized Linear Mixed Models

A.A. Sunethra[1],
Lecturer of Department of Statistics,
University of Colombo, Sri Lanka

M. R. Sooriyarachchi[2]
Professor of Department of Statistics,
University of Colombo, Sri Lanka

Abstract— **The literature clearly demonstrated how the random effect miss-specification in Generalized Linear Mixed Models (GLMMs) affect the model performance with respect to the Type II Errors of the Type III F-test. The method of Sandwich Variance Estimation (SVE) is a very popular method for improving the functionality of miss-specified models. This study attempted on examining whether the use of SVE could improve the Type II Errors of miss-specified GLMMs. A comprehensive simulation study comprising data from a Binary Logistic Mixed Model was performed of which the results clearly demonstrated that Type II Errors are being affected by random effect miss-specification. The novel finding of the study was that the adoption of SVE failed to contribute significantly to improve the functionality of GLMMs when random effects of the GLMMs are not correctly specified.**

Keywords – Generalized Linear Mixed Models, Sandwich Varaince Estimation, Random Effect Miss-specification, Binary Logistics Mixed Model,

I. INTRODUCTION

Even a quick glance on theories behind many statistical terms, tests, procedures and models would reveal that they are defined for independent and identically distributed data. As [1] has mentioned, "the first line in the description of many common statistical tests is that the data represent independent samples from the same statistical distribution; that is the sampled observations must be identically and independently distributed (iid)". But, quite contrarily, there are many data scenarios that violate the independence assumption between the observations in the data. Examples include repeated measures data, longitudinal data, hierarchical data anthese kinds of data are termed as clustered/correlated data in statistical terminology. As [2] had explained "Clustered data arise when the data from the whole study can be classified into a number of different groups, referred to as clusters".

The examples in the literature that had focused on correlated/clustered data analysis mainly falls under four distinct approaches; namely (i). ignoring clustering, (ii). reducing clusters to independent observations, (iii) fixed effects regression/ANOVA approaches, and (iv) explicitly accounting for clustering [2]. The method of robust variance estimation which is often nicknamed as Sandwich Variance Estimation (SVE) has been a popular method for analysis of correlated data which was proposed initially by [3] as a method for improving the standard errors of the maximum likelihood estimators of miss-specified models. The use SVE for analyzing correlated data was in such a way that models assuming independency was fitted for correlated data which make the models to be miss-specified and the standard errors of the models were estimated using SVE as it is intended for improving the standard errors of miss-specified models. This approach was of high appreciation when specialized statistical models were not available for correlated data. But, later with the development of specialized statistical models for clustered data, the necessity and impact of using SVE in such models was on argument among the researches.

The class of Mixed Models [4] is a statistical modeling approach developed for correlated/clustered data analysis where Generalized Linear Mixed Models (GLMMs) are an extension of generalized linear models for the analysis of non-Gaussian correlated data. In GLMMs, the presence of clusters in the data is introduced to the model as a random effect which follows a certain probability distribution. For example, when the data is clustered on a clinic/geographical area, the clinic/geographical area is introduced to the model as a random effect which follows a probability distribution which is mostly assumed to be normally distributed. Therefore, one major assumption in GLMMs is the choice of the probability distribution made for the random effect. Most of the statistical packages which facilitate fitting GLMMs also include only the Normal distribution as the choice of the random effect distribution. It is noteworthy that these random effects are not observable and hence it is difficult / impossible to validate the assumption of the probability distribution assumed for random effects. Many researches in the literature have demonstrated that miss-specifications associated with GLMMs are mostly due to the errors made with the distributional assumption for the random effects [5] [6] [7]. But, [8] have shown theoretically that the maximum likelihood estimators obtained for linear mixed models are consistent and asymptotically normally distributed even though the random effects are miss-specified. But, authors have showed that this property is not

DOI : 10.5176/2251-3388-3.2.64

held within Generalized Linear Mixed Models [4] [5]. Therefore, miss-specification of random effects in GLMMs makes an impact on the performance of GLMMs. As mentioned earlier, the method of SVE is meant for improving standard error estimation of miss-specified models, it was appealing to examine whether Sandwich Variance Estimation can impact on miss-specified Generalized Linear Mixed Models. Therefore, the objective of this study was to examine comprehensively how the SVE can impact on miss-specified GLMMs where the miss-specification is caused by the random effect miss-specification.

II. Literature

A thorough literature review was undertaken on miss-specified GLMMs , particularly on finding how SVE is used in GLMMs for confronting miss-specification. Many authors [6] [10] [5] have written on the impact of miss-specified GLMMs considering the properties of parameter estimates in miss-specified GLMMs. Reference [5] have shown how Type I and type II errors of Wald test are affected by the random effect miss-specification using a simulation study consisting of Binary repeated measures data. Their simulated data consisted of random effects from various distributions like Normal, Power function, Mixture of Normal and etc and the data were analyzed by a GLMM fitted assuming normal distribution for the random effects. Then, the Type I and Type II errors of the Wald test which tests the significance of fixed effect parameters were compared to envisage how the choice of the random effect distribution impact on the GLMM. It was revealed that Type I error could be maintained within the desired 95% probability interval whereas the Type II errors were severely affected by the miss-specification. But when the variable considered for the Wald test is included in the random effect structure (for Eg. Intercept parameter), even the type I errors were severely inflated [5]. Later on, [10] highlighted some pitfalls with the design of the simulation study used by [5] which led [11] to perform another simulation study on which they considered Type II error under random effects misspecification in Generalized Linear Mixed Models. They concluded their findings as "misspecification of the random effects distribution in GLMM can have an effect on the ML estimators and inferential procedures". So, their main concern was to investigate whether random effect miss-specification affect the functionality of GLMMs for which they have found out that different aspects of the model are affected in different ways and to different degrees while this impact seemed to depend on the complexity of the random-effects structure, the variance of the underlying random-effects distribution, and the parameters of interest. In summary, these examples in literature have stressed on probable miss-specifications of GLMMs and how they affect the functionality of GLMMs. But, none of these researches attempted to examine the role of SVE to confront miss-specifications in GLMMs. Reference [12] demonstrated the use of SVE for analyzing Binary repeated measures data using GLMMS. But, the design of the simulation study used by did not clearly distinguish what miss-specifications were present in the data with respect to the assumptions of GLMMs though they showed SVE improved the miss-specified GLMMs fitted for the simulated data. In contrast, this study addresses specifically random effect miss-specification of GLMMS which is the most commonly encountered misspecification in GLMMs as per literature highlights and known to impact on the performance of GLMMs as well. So, this study fills the gap in the literature by examining whether Sandwich Variance Estimation can make a significant contribution to improve the functionality of GLMMs which is affected by the random effect misspecification particularly on improving the Type II Errors of the fixed effect parameters in miss-specified GLMMs.

III. Methods

To meet the intended objective of the study, a simulation study was performed for which the design of the simulation study used by [5] [10] and [11]. It is noteworthy, this design of the simulation study is been extensively used by these authors to identify random effect miss-specification of GLMMs and hence of great relevance to the current study as well.

The data were simulated on a clustered Binary responses from a logistics random intercept model given below.

$$\text{Logit}(P(y_{ij}=1|b_i)) = \beta_0 + \beta_1 Z_i + \beta_2 t_j + b_i \quad (1)$$

where

y_{ij}= binary response of the i^{th} subject at j^{th} time point

Z_i= 1(0) denotes the treatment (control) group of the i^{th} subject

t_j= denotes the occasion of measurement with values 0, 1, 2, 4, 6 and 8.

b_i= denotes a random intercept $b_i \sim G$

Following the strategy of [5] [10] and [11], it was set β_0=-8 β_2=1 and three different values for the treatment effect (β_1) were considered such as β_1=0.5, 1 and 5.

Reference [5] initially considered four random efffect distributions namely; Normal, Power Function, Discrete and Mixture of Normal and they found out that a significant discrepancy was held between Power function distribution and Normal distribution which led them to consider only the Power and Normal distributions in their subsequent works [10] [11]. Hence, this study also used only Power function distribution for the true random effect distribution while models were fitted assuming as if random effects are Normally distributed. The density of the Power function distribution G ~Power(¤,©) takes the form:

$$g(b) = \frac{\gamma b^{\gamma-1}}{\theta^{\gamma}} \quad (2)$$

It was set γ=80 and the value of θ was varied according to the variance of the random effect distribution considered. As mentioned by [5], only type II errors are affected by random effect misspecification, hence this study also used only the type II errors as the measurement for evaluating whether SVE has made significant improvement in Type II errors or not.

Therefore, the case of $\beta_1=0$ was not considered here, since the rejection of the Type III F-test when $\beta_1=0$ corresponds to Type I error and [5] has found out that Type I errors are not affected by the random effect miss-specification on the binary logistic model considered here particularly when the covariate of interest is not included in the random effect structure. Three different sample sizes were used, namely 25,100 and 400 with 500 replicates of each sample size. The number of rejections of the Wald test corresponding to H_0: $\beta_1=0$ was considered which gives rise to the Type II errors/Power.

The data were simulated by writing an R program along the lines of [5] while SAS procedure Proc GLIMMIX was used to fit the models. The method of parameter estimation used was Adaptive Gaussian Quadrature with 50 cut points as per suggestions of [5] and [6]. Two method of variance estimation was used when fitting the Binary logistic mixed model, namely model based variance estimator and Sandwich Variance Estimator. A comparison of the Type II errors/Power was made between two types ofGLMMs.

IV. Results

Following table gives the Type II errors attained by the two type of GLMMS (default GLMM and GLMM with SVE) for Wald test for H_0: $\beta_1=0$ with respect to value of the β_1 (0.5, 1, 5) and with respect to the variance of the random effect distribution (1,4,16).

Table 1 : Type II Errors of the Type III F-test

Beta β_1	Variance	Sample Size	Type II error	
			Default Variance	SVE
0.5	1	25	0.08	0.11
		100	0.27	0.27
		400	.8	0.79
	4	25	0.05	0.07
		100	0.2	0.2
		400	0.61	0.62
	16	25	0.04	0.06
		100	0.13	0.13
		400	0.39	0.39
1	1	25	0.21	0.23
		100	0.78	0.78
		400	1	1
	4	25	0.12	0.17
		100	0.58	0.578
		400	1	1
	16	25	0.08	0.11
		100	0.32	0.33
		400	1	1
5	1	25	.69	0.73
		100	1	1
		400	1	1
	4	25	0.89	.9
		100	1	1
		400	1	1
	16	25	0.85	0.89
		100	1	1
		400	1	1

The results of the simulation study tabulated above have shown that irrespective of the value assumed for β_1 and irrespective of the varaince of the random effects, Type II Error of the type III F-test is very low at small sample of size 25. At large sample of size 400, satisfcatory acheivement is been made in Type II Errors at low level of varaince (i.e varaince =1) in the random effects, but when the varaince of the random effect is high, power was low even at large sample of size 400. So, these findings clealy higlighted that Power/Type II Errors of the Type III F-test is affected by the random effect miss-specification. The important and the novel finding of this study is that the in-depth simuation study performed here indicated with evdence that the adoption of the classicle SVE has not been able to improve signicantly the power of the Type III F-test at any setting. At $\beta_1 = 5$, Power results appeared to be satisfcatory even at small sample sizes and at high levels of varaince in random effects which might be due to cause that the assumed value for β_1 is largely different from 0 which is the hypothised value of the Type III F-test. Thus, these results indicated that the adoption of Sandwich Variance Estimation in GLMMS is unable to improve Type II Errors of the Type III F-test caused by the random effect miss-specification of GLMMs. In genral, the results indicated that a satifactory Type II errors are held at

large sample sizes while at small sample sizes a Type IIs have dropped significantly. Since classical SVE was unable to imrpove this miss-specification, small sample versions of the SVE were als appkied. But, these modified Sandwich Varaince Estimators resulted non-convergent modelt fits.

V. Discussion

The main counterpart of this study was to examine whether SVE can enhance the functionality of miss-specified GLMMS where the miss-specification is caused by random effect miis-specification. The results of the study revealed that such an improvement is not attainable though the use of SVE. Though reference [12] has shown up that SVE can improve miss-specified GLMMs, it was not identified what miss-specification of GLMMs can be improved/not improved by the use of SVE in GLMMs. Thus, the findings of this study highlighted with evidence that the miss-specifications caused by violating the assumption of the random effect distribution could not be overcome by using Sandwich Variance Estimation as the method of variance estimation in GLMMs.

Acknowledgment

The financial support given by the University of Colombo, Sri Lanka by awarding a Postgraduate research scholarship to conduct research work is highly appreciated.

References

1. Ives , A., & Jun Zhu . (2006). Statistics For Correlated Data: Phylogenies, Space, And Time. Ecological Applications, 20-32.

2. Sally, Galbraith; James A. , Daniel; Bryce , Vissel. (2010). A Study of Clustered Data and Approaches to Its Analysis. The Journal of Neuroscience, • 30(32):, 10601–10608.

3. Huber, P. J. (1967). The behavior of maximum likelihood estimates under non-standard conditions. . In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (Vol.1, pp. 221–233). Berkeley, CA: University ofCalifornia

4. Agresti, A. (2002). Categorical data analysis. N.J. Wiley.

5. Litiere, S., Alonso , A., & Molenbergh, G. (2007). Type I and type II error under random-effects misspecification in generalized linear mixed models. Biometrics, Vol. 63, No. 4.

6. Litière, S., Alonso, A., & Molenberghs, G. (2008). The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. Statistics in Medicine, 3125-3144.

7. Chavance, M., & Escolano , S. (2012). Misspecification of the covariance structure in generalized linear mixed models. Statistical methods in medical research.

8. Verbeke,G., & Lesaffre, E.(1997). The effect ofmisspecifying the random-effects distribution in linear mixed models for longitudinal data. Computational Statistics & Data Analysis, 541-556.

9. Huber, P. J. (1967). The behavior of maximum likelihood estimates under non-standard conditions. . *In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (Vol.1, pp. 221–233). Berkeley, CA: University of California Press*

10. Neuhaus, J., McCulloch, C., & B, R. (2011, Jun). A note on Type II error under random effects misspecification in generalized linear mixed models. *Biometrics*, 654-66

11. Litiere, S., Alonso, A., & Molenberg, G. (2011). Rejoinder to "A note on Type II error under random effects misspecification in Generalized Linear Mixed Models". Biometrics, 67(no.2),656-660.

12. Sunethra, A.A, & Sooriyarachchi, M.R. (2016). Use of Sandwich Variance Estimation in Generalized Linear Mixed Models: for Binary Repeated Mesures Data. In C. B. Guptal (Ed.), 4th Annual International Conference on Operations Research and Statistics. Global Science and Technology Forum(GSTF).

A.A. Sunethra is a lecturer from the Department of Statistics, University of Colombo and is currently reading for her Ph.D. Her research interests include survival analysis, joint modeling and R programming.

Roshini Sooriyarachchi, is a professor in the Department of Statistics, University of Colombo. She has a bachelor's degree (first class) in Physical Science and a postgraduate diploma in Applied Statistics from the University of Colombo, Sri Lanka and a M.Sc. in Biometry and a Ph.D. in Applied Statistics from the University of Reading, UK. Her area of specialization is Medical Statistics. She has nearly 100 publications in peer-reviewed journals and conference proceedings with nearly 200 citations to herarticles.