International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:13, No:4, 2019

Comparison of Methods of Estimation for Use in Goodness of Fit Tests for Binary Multilevel Models

I. V. Pinto, M. R. Sooriyarachchi

***Abstract*—**It can be frequently observed that the data arising in our environment have a hierarchical or a nested structure attached with the data. Multilevel modelling is a modern approach to handle this kind of data. When multilevel modelling is combined with a binary response, the estimation methods get complex in nature and the usual techniques are derived from quasi-likelihood method. The estimation methods which are compared in this study are, marginal quasi-likelihood (order 1 & order 2) (MQL1, MQL2) and penalized quasi-likelihood (order 1 & order 2) (PQL1, PQL2). A statistical model is of no use if it does not reflect the given dataset. Therefore, checking the adequacy of the fitted model through a goodness-of-fit (GOF) test is an essential stage in any modelling procedure. However, prior to usage, it is also equally important to confirm that the GOF test performs well and is suitable for the given model. This study assesses the suitability of the GOF test developed for binary response multilevel models with respect to the method used in model estimation. An extensive set of simulations was conducted using MLwiN (v 2.19) with varying number of clusters, cluster sizes and intra cluster correlations. The test maintained the desirable Type-I error for models estimated using PQL2 and it failed for almost all the combinations of MQL. Power of the test was adequate for most of the combinations in all estimation methods except MQL1. Moreover, models were fitted using the four methods to a real-life dataset and performance of the test was compared for each model.

***Keywords*—**Goodness-of-fit test, marginal quasi-likelihood, multilevel modelling, type-I error, penalized quasi-likelihood, power, quasi-likelihood.

I. Introduction

FREQUENTLY data structures can be observed with a hierarchical or a nested structure attached with the data. These data arise from various fields such as, medical field where patients are nested within hospitals, educational field where students are nested within schools etc. More generally, this is referred to as "units" nested at different "levels" of the hierarchy. Multilevel data are the data structures which consist of two or more levels. For example, children are level-1 units and schools are level-2 units and children from the same school may tend to behave in a similar manner than children from other schools. Even though multilevel data can have more than two levels, similar to the examples mentioned above, this study is only based upon the two-level multilevel structure.

A. Multilevel Modelling

Approaches used in the past tend to ignore the existence of a hierarchy in multilevel data. These methods treat all the units at its lowest level and conduct standard analysis techniques. However, then it violates the assumption of independence of observations which is an important assumption in standard statistical techniques. There are several methods introduced to cater to this problem in multilevel data [3]. Among such methods, this paper is conducted upon multilevel modelling (MLM) technique.

MLM can be categorized based on the distribution of the response variable, type of data structure and the variance structure [4] Considering the distribution of the response variable, this study considers the binary response variable with the logit model. The data structure considered in the study is the simplest and the most common data structure, two-level hierarchical structure. Considering the variance structure, it is based on the random intercept model where only the intercept is allowed to vary randomly. Thus, the model used throughout is the random intercept, binary logistic MLM.

B. Methods of Estimation

There are several methods for estimation of parameters in MLM. Reference [2] has mentioned several estimation procedures such as maximum likelihood method, generalized least squares method and generalized estimating equations. Also, there are Bayesian methods such as Markov Chain Monte Carlo (MCMC). However, binary logistic MLM leads to more complex estimation procedures. The most commonly used approach is to approximate the nonlinear link function by using a nearly linear link and include the effects of MLM. This approach is called the quasi-likelihood approach [2]. Following this approach, this study is based upon four estimation procedures namely;

1. Marginal Quasi Likelihood – Order 1 (MQL1)
2. Marginal Quasi Likelihood – Order 2 (MQL2)
3. Penalized Quasi Likelihood – Order 1 (PQL1)
4. Penalized Quasi Likelihood – Order 2 (PQL2)

The fascinating fact with these methods is that there is no one best method. References [1] and [5] have found out the behavior of these methods with varying multilevel structures. However, the impact of these methods of estimations on the performance of GOF tests is a novel research area.

C. GOF Tests

Statistical models are of no use if they provide the user with misleading results. If the models do not fit the data but are blindly used, the results will be erroneous and might lead to biased conclusions. Therefore, similar to model fitting, checking the GOF of the model is also an important step. It is equally important to make sure that the GOF test performs

I. V. Pinto is with University of Colombo, Sri Lanka (phone: 94-779995185; e-mail: vimukthinipinto@gmail.com)
M. R. Sooriyarachchi, Senior Professor is with the University of Colombo, Sri Lanka (e-mail: roshini@stat.cmb.ac.lk).

well and the test does not recommend incorrect models. Therefore, it is essential to use a test which is proved to have a good performance and which is applicable to the model under consideration.

Due to the nature of the response, the procedures to conduct tests are different between continuous and categorical models. Moreover, due to the correlations that exist between observations, GOF tests for single level models are not suitable for MLM. By taking the basis from Hosmer and Lemeshow test [6] for single level logistic models and test developed by Lipsitz et al. [7] for single level ordinal logistic models, Perera et al. [8] proposed a GOF test for multilevel binary logistic models by marking a remarkable advancement in the field of MLM. However, the GOF of the test was only examined for models estimated using PQL2.

D. Objective of the Study

One can conduct estimations in MLM by using any estimation method. However, the developed GOF test might not suit all these methods. The main objective of the study is to recommend the estimation procedures where the GOF test is applicable under different multilevel structures. To identify the applicability of the test with the method of estimation, the study aims to compare the performance of the test in terms of type-I error and power of the test under each estimation method using simulations. Moreover, to conduct comparisons practically, models are fitted using the four methods to Bangladesh fertility survey (1989), an in-build dataset in MLwiN and the GOF test is assessed.

The methodology used in the study is explained under 'Methodology' section while simulation results of the study along with conclusions are presented under 'Results and Discussion' followed by 'Conclusions'.

II. Methodology

Presented below is the random intercept model with single explanatory variable x_{ij}, where $i=1,2,\ldots,n_j$ and $j=1,2,\ldots,k$, where k is the number of clusters.

$$logit\ (\pi_{ij}) = \beta_{0j} + \beta_1 x_{ij} \quad (1)$$

where $\beta_{0j} = \beta_0 + u_{0j}$ and $u_{0j} \sim N(0,\sigma_{u0}^2)$

The alternative model proposed by [8] with the indicator variables is as:

$$logit\ (\pi_{ij}) = \beta_{0j} + \beta_1 x_{ij} + \sum_{g=2\ to\ 10}(\gamma_g I_{gij}) \quad (2)$$

where $\beta_{0j} = \beta_0 + u_{0j}$ and $u_{0j} \sim N(0,\sigma_{u0}^2)$

$$\sum_{g=2\ to\ 10}(\gamma_g I_{gij}) = \gamma_2 I_{2ij} + \gamma_3 I_{3ij} + \ldots + \gamma_{10} I_{10ij}$$

where $i=1,2,\ldots,n_j$ and $j=1,2,\ldots,k$, where k is the number of clusters.

A. Simulation Study

To simulate the single explanatory variable, [9] has suggested using Bernoulli distribution, normal distribution or uniform distribution. Hence, moving alongside with the original proposer of the test [8], the explanatory variable is simulated from a normal distribution with mean of 2.0 and standard deviation of 1.0.

Given above are the two types of models fitted under the simulations study. These models are fitted under various conditions and the GOF of model (1) is assessed based on the Type-I error and power of the test by using the joint Wald statistic with the following hypothesis of interest.

- H_0: $\gamma_2 = \gamma_3 = \ldots = \gamma_{10} = 0$
- H_a: At least one coefficient of the indicator variables is not zero

Then the joint Wald statistic obtained from MLwiN is compared with the chi- square value of 9 degrees of freedom at 5% significance level (16.919). Various combinations considered in the study are summarized in Table I.

TABLE I
Factors Considered in the Simulation

Factor	Number of Combinations	Values
Standard deviation of the random component	3	1, 1.5, 3
Number of clusters	2	15, 60
Number of observations in each cluster (Cluster size)	2	20, 50
Method of approximation	2	MQL, PQL
Order of the Taylor series	2	1, 2
Total Combinations	$3 \times 2 \times 2 \times 2 \times 2 =$ **48**	

The standard deviations of the random component reflect the intra cluster correlations (ICC) and these values are selected to be in the desirable ranges of ICC [10]. Moreover, the number of clusters present in the model and the number of observations per each specified cluster are selected according to the specified guidelines [11], [12].

B. Real-Life Application

To compare the practical results obtained by the GOF test with the change in estimation procedure; models are fitted using the four methods of estimations. Next, the GOF test is applied to the models to compare the recommendations given by the test under each method. The inbuilt dataset used, a sub sample of '1989 Bangladesh fertility survey dataset' [13] consists of 2687 records of data collected from Bangladesh women over the country nested within their district of residence. The dataset comprises of a binary response variable, 'use' which indicates whether the individual use contraceptives at the time of data collection or not. Table X provides a summary of the variables considered in the study.

Districts in the dataset are coded from 1 to 61 where no observations are reported from district 54. Thus, a total of 60 districts are available in the dataset. However, number of women belonging to some districts was less than 10. Thus, to avoid any complications in applying the GOF test with the use of 10 indicator variables, these districts are ignored in the model fitting. Moreover, districts which contain women between 10 and 20 are also ignored to avoid any possibilities of non-convergence. An R code is used for easy removal of

these districts from the dataset. Thus, final dataset contained 2711 women in 49 districts. The model which is suitable to this kind of data is the binary logistic MLM. In order to identify the most important variables and build up the final model, forward selection is implemented with the use of Wald statistic at 5% level of significance for models estimated using each method. Then, the GOF test is applied on the best models developed from forward selection.

To apply the GOF test, under each estimation method, probabilities are predicted and these probabilities are sorted in ascending order within each district. Each district is divided in to 10 groups such that closer probabilities belong to one group. Unlike in the simulation study, most of the districts in the dataset are not divisible by 10. Thus, according to the suggestion made by [14], indicator variables are defined as,

$$g_j = (\text{Number of observations in the } j^{th} \text{ cluster})/10$$
$$\text{If } i \leq a \times g_j \text{ then } I=a$$

where $a=1,2,\ldots,10$; $j=1,2,\ldots,49$ and i is the i^{th} observation in the j^{th} cluster which is sorted in ascending order. An R code is used in inserting these new indicator variables to conduct the GOF test and models are fitted using MLwiN.

III. Analysis, Results and Discussion

A. Simulation Study

To assess the performance of the test, Type-I error and power of the test is assessed under the combinations listed under Table I.

1) Type-I Error Results

Type I error occurs as a result of rejecting the null hypothesis which is actually true. The probability of making a Type-I error is denoted using the Greek symbol α. To lower the risk of Type I error, α is set to the value 0.05. After generating data under the correct null hypothesis for 1000 datasets, the number of times it rejects the null hypothesis is obtained and it is checked if it is within the 95% probability interval for α (0.036,0.064). Type-I error rates obtained for each estimation method are shown in Tables II-V.

TABLE II
Type-I Error Results for MQL1

ICC	No: of Clusters	Cluster Size	Runtime	Significant Datasets	Rej Prop	Results
ICC 1	15	20	1m40s	5 from 387	0.013	Outside the limit
	15	50	1m50s	14	0.014	Outside the limit
	60	20	1m54s	11	0.011	Outside the limit
	60	50	2m12s	13	0.013	Outside the limit
ICC 2	15	20	1m15s	7	0.007	Outside the limit
	15	50	1m14s	8	0.008	Outside the limit
	60	20	1m48s	0	0	Outside the limit
	60	50	2m34s	5	0.005	Outside the limit
ICC 3	15	20	1m17s	4	0.004	Outside the limit
	15	50	1m17s	2	0.002	Outside the limit
	60	20	1m37s	2	0.002	Outside the limit
	60	50	2m02s	2	0.002	Outside the limit

- When the ICC is small in the smallest sample size, there are convergence issues.
- None of the combinations produced a desirable Type I error within the acceptable region with the use of MQL1 method. Interestingly, all these proportions produced are below the lower bound.
- For a given cluster size, it seems that the rejection proportion decreases with the increase in ICC.

TABLE III
Type-I Error Results for MQL2

ICC	No: of Clusters	Cluster Size	Runtime	Significant Datasets	Rej Prop	Results
ICC 1	15	20	2m 10s	51	0.051	Within the limit
	15	50	1m52s	42	0.042	Within the limit
	60	20	2m28s	65	0.065	Outside the limit
	60	50	3m12s	65	0.065	Outside the limit
ICC 2	15	20	2m11s	7 from 414	0.017	Outside the limit
	15	50	1m17s	10 from 238	0.042	Within the limit
	60	20	2m38s	65	0.065	Outside the limit
	60	50	3m05s	86	0.086	Outside the limit
ICC 3	15	20	1m28s	5 from 422	0.012	Outside the limit
	15	50	2m13s	18	0.018	Outside the limit
	60	20	3m05s	82	0.082	Outside the limit
	60	50	4m13s	132	0.132	Outside the limit

- More Type-I error convergence issues are present with the models estimated using MQL2.
- Similar to MQL 1, for a given cluster size, it seems that the rejection proportion decreases with the increase in ICC.
- Except combinations with standard deviation 1.0 (ICC 1), Type-I error rate of the test seems to increase with the increase in sample size.

TABLE IV
Type-I Error Results for PQL1

ICC	No: of Clusters	Cluster Size	Runtime	Significant Datasets	Rej Prop	Results
ICC 1	15	20	2m02s	12 from 387	0.031	Outside the limit
	15	50	2m10s	30	0.03	Outside the limit
	60	20	2m49s	29	0.029	Outside the limit
	60	50	2m23s	38	0.038	Within the limit
ICC 2	15	20	2m03s	32	0.032	Outside the limit
	15	50	2m16s	23	0.023	Outside the limit
	60	20	3m05s	29	0.029	Outside the limit
	60	50	2m50s	36	0.036	Within the limit
ICC 3	15	20	2m28s	32	0.032	Outside the limit
	15	50	2m28s	38	0.038	Within the limit
	60	20	3m17s	24	0.024	Outside the limit
	60	50	4m09s	36	0.036	Just within the limit

- Most of the combinations produce undesirable Type-I errors with PQL1 method. However, Type-I error holds for four of the combinations.
- Only 387 datasets converged for the smallest sample size with the smallest ICC, and only 5 datasets resulted in higher Wald statistics. This is the same situation as in MQL 1.
- Unlike previous two methods, there is no visible pattern in Type I errors with the increase in standard deviations

for a given sample size.

TABLE V
TYPE-I ERROR RESULTS FOR PQL2

ICC	No: of Clusters	Cluster Size	Runtime	Significant Datasets	Rej Prop	Results
ICC 1	15	20	2m40s	51	0.051	Within the limit
	15	50	2m55s	42	0.042	Within the limit
	60	20	3m33s	47	0.047	Within the limit
	60	50	4m42s	45	0.045	Within the limit
ICC 2	15	20	2m05s	39	0.039	Within the limit
	15	50	2m58s	38	0.038	Within the limit
	60	20	4m23s	42	0.042	Within the limit
	60	50	3m20s	42	0.042	Within the limit
ICC 3	15	20	3m41s	41	0.041	Within the limit
	15	50	3m55s	47	0.047	Within the limit
	60	20	4m38s	33	0.033	Outside the limit
	60	50	4m09s	36	0.036	Just within the limit

- Type-I error of the test is within the desirable range for all the combinations except for one combination. Even for the combination which does not produce a Type-I error within the range, it is only lower by 0.003 units from the lower bound.
- There seems no pattern in Type-I errors of the test when estimations are done in PQL2.

2) Power Results

Power of a test is associated with the type-II error (β). It is the probability of rejecting the null hypothesis when the alternative hypothesis is true. The β is allowed to vary and it is desirable to have a lower β. Thus, power is also allowed to vary and the high values of power are considered to be better. Unlike the simulations under type I error, two sets (1000 each) of datasets are simulated by considering the explanatory variable.

- Set 1: 1000 datasets are simulated such that xij~ N(2,1)
- Set 2: 1000 datasets are simulated such that xij ~ N(2,4)

TABLE VI
POWER RESULTS FOR MQL1

ICC	*No: of Clusters*	*Cluster Size*	x_{ij} ~ N(2,1)		x_{ij} ~ N(2,4)	
			Runtime	*Rej Prop*	*Runtime*	*Rej Prop*
ICC 1	15	20	2m12s	0.027	1m40s	0.103
	15	50	1m22s	0.06	1m17s	0.66
	60	20	1m31s	0.086	1m42s	0.688
	60	50	1m55s	0.445	2m11s	1
ICC 2	15	20	1m07s	0.009	1m12s	0.043
	15	50	1m11s	0.025	1m19s	0.473
	60	20	1m33s	0.043	1m39s	0.482
	60	50	2m13s	0.266	2m04s	0.999
ICC 3	15	20	1m19s	0.002	1m31s	0.02
	15	50	1m20s	0.011	1m27s	0.299
	60	20	1m37s	0.012	1m43s	0.277
	60	50	1m53s	0.121	2m15s	0.989

The second set is generated to improve the power values which are generated from first set. As [8] pointed out, when the random effect is larger than the covariate effect, the explanatory power of the explanatory variable is reduced. Thus, the 2nd set is simulated with a variance of 4.0. To evaluate the power associated with the simulation, data are generated from the false null hypothesis using an MLwiN Macro code and the proportion of rejecting the false null hypothesis (power of the test) is calculated.

- There seems a clear improvement in the power of the test when the explanatory variable is generated with a larger variance.
- When the standard deviation of the random component increases, power of the test clearly decreases.
- When sample size increases there seems a clear increment in power of the test.
- According to the 80% rule set by [15], when models are fitted using MQL1, only three combinations (combinations with largest sample size) produce power values more than 80%.

TABLE VII
POWER RESULTS FOR MQL2

ICC	*No: of Clusters*	*Cluster Size*	x_{ij} ~ N(2,1)		x_{ij} ~ N(2,4)	
			Runtime	*Rej Prop*	*Runtime*	*Rej Prop*
ICC 1	15	20	1m48s	0.071	1m33s	0.202
	15	50	5m12s	0.183	1m44s	0.803
	60	20	2m05s	0.26	2m34s	0.873
	60	50	2m33s	0.683	3m05s	1
ICC 2	15	20	2m540s	0.057	1m10s	0.132
	15	50	2m42s	0.132	2m01s	0.688
	60	20	2m41s	0.292	2m34s	0.857
	60	50	3m44s	0.681	3m19s	1
ICC 3	15	20	40s	0.016	1m00s	0.059
	15	50	2m18s	0.084	2m29s	0.518
	60	20	2m53s	0.303	3m05s	0.827
	60	50	3m22s	0.633	3m40s	1

TABLE VIII
POWER RESULTS FOR PQL1

ICC	*No: of Clusters*	*Cluster Size*	x_{ij} ~ N(2,1)		x_{ij} ~ N(2,4)	
			Runtime	*Rej Prop*	*Runtime*	*Rej Prop*
ICC 1	15	20	1m47s	0.051	2m24s	0.168
	15	50	1m56s	0.127	2m29s	0.771
	60	20	2m19s	0.153	2m48s	0.788
	60	50	2m58s	0.609	3m29s	1
ICC 2	15	20	3m07s	0.051	2m17s	0.15
	15	50	3m12s	0.12	2m27s	0.718
	60	20	3m10s	0.135	3m11s	0.755
	60	50	4m04s	0.558	3m48s	1
ICC 3	15	20	2m38s	0.046	3m11s	0.122
	15	50	2m44s	0.116	3m03s	0.672
	60	20	3m08s	0.133	3m53s	0.659
	60	50	4m02s	0.514	4m49s	1

- The first three points listed under MQL1 with Table VI remains consistent with the finding for MQL2 as well.
- Unfortunately, many convergence issues are present when models are fitted using MQL2. Such combinations are shown in red.
- Considering the 80% rule, test produces high power values for all the combinations with large number of

clusters (60).

Power results for models fitted using PQL1 behaves similar to MQL1 models. Thus, the four points listed for MQL1 holds for PQL1 as well. However, GOF test produces considerably high-power values here.

TABLE IX
POWER RESULTS FOR PQL2

ICC	No: of Clusters	Cluster Size	$x_{ij} \sim N(2,1)$		$x_{ij} \sim N(2,4)$	
			Runtime	Rej Prop	Runtime	Rej Prop
ICC 1	15	20	2m26s	0.071	2m48s	0.203
	15	50	2m05s	0.161	3m17s	0.803
	60	20	3m11s	0.203	3m41s	0.844
	60	50	4m04s	0.631	4m45s	1
ICC 2	15	20	2m53s	0.066	3m03s	0.194
	15	50	2m57s	0.158	3m11s	0.756
	60	20	4m36s	0.179	5m03s	0.794
	60	50	6m02s	0.589	6m12s	1
ICC 3	15	20	3m52s	0.063	4m23s	0.164
	15	50	4m00s	0.153	4m33s	0.709
	60	20	4m22s	0.173	5m30s	0.726
	60	50	5m23s	0.54	7m42s	1

- Power of the test for the models fitted using PQL2 is also consistent with the first three points under MQL1 method.
- No convergence issues are encountered.
- Considering the 80% rule, GOF test produces considerably high number of combinations, more than 80% when models are fitted using PQL2. Even though 80% is not reached, the test produces a power of at least 70% for all the combinations except the lowest sample size. Reference [16] explains on the common situation of low power when the sample size is less than ideal.

B. Real-Life Application

TABLE X
DESCRIPTION OF VARIABLES

Variable	Type	Description
Woman	Factor	Code for each woman (Level 1): 2687 women
District	Factor	Code for each district (Level 2): 60 districts
Use	Binary	Use of contraceptive at the time of the survey **1**: Using contraception **0**: Not using contraception
Lc	Categorical (Ordinal)	Number of living children at the time of the survey **0**: No children (Reference) **1**: 1 child **2**: 2 children **3**: 3 or more children
Age	Continuous	Age in years of the person at the time of the survey. This is centered to 30 years.
Urban	Binary	The type of region of residence **0**: Rural (Reference) **1**: Urban
Educ	Categorical (Ordinal)	The level of education of the person **1**: None (Reference) **2**: Lower Primary **3**: Upper primary **4**: Secondary or above
Hindu	Binary	Religion of the individual **0**: Muslim (Reference) **1**: Hindu
d lit	Continuous	Proportion of literate women in the district
d pray	Continuous	Proportion of Muslim women in the district who pray every day

The most important variables out of the variables presented in Table X are selected using the forward selection procedure as explained in the 'Methodology'.

All the four best models resulting from forward selection contain the same variables (Variables except *d_lit*). However, parameter values and their standard errors are different for each estimation method. The GOF test is individually applied to the four models and following are the joint Wald statistics obtained.

TABLE XI
TEST APPLICATION

Estimation Method	Joint Wald Statistic	p-Value
MQL1	8.535	0.48125
MQL2	8.838	0.45236
PQL1	8.340	0.50028
PQL2	8.643	0.47086

Comparing the joint Wald statistics obtained with the critical value, chi-square value at 5% significance level, (16.919) the four models provide lower joint Wald statistics, implying that the models fit the data well. A simulation is conducted to approximate the type I error and power of the test for each method of estimation before making recommendations from the test. To closely match the dataset which comprises of 2711 units with 49 districts, a balanced dataset is simulated with a sample size of 2700 consisting of 45 clusters by taking coefficients to closely match the real models. Simulations provided the following results.

TABLE XII
SIMULATION TO MATCH THE REAL-LIFE DATASET

Estimation Method	Type-I Error	Power
MQL1	0.036	0.975
MQL2	0.057	0.978
PQL1	0.050	0.979
PQL2	0.052	0.978

Interestingly, results obtained from all four methods produce type I errors within the boundary (0.036, 0.064) and satisfactory powers from the test. However, α for MQL1 is more towards the lower margin (just within the limits). For this dataset, the ideal α of 0.05 and the highest power is produced by PQL1.

IV. CONCLUSION

The conclusions which could be derived from the study are,

- The GOF test performs differently with the models estimated using the four methods of estimation.

Conclusions on Type-I error of the test given by the simulation study are as follows.

- The Type I-error simulations indicated that the test produces adequate Type-I errors for models estimated using PQL2.
- The test fails to maintain the Type-I error for models estimated using MQL1.
- For models estimated using MQL2 or PQL2, the test seems to produce adequate Type-I errors depending on

the sample size.

Power of the test depends on the selected incorrect functional form. Following conclusions could be derived from the simulations conducted using the specified functional form discussed earlier.

- Power of the test increases with the increase in sample size irrespective of the method of estimation.
- There seems an inverse relationship between power of the test and ICC.
- The GOF test produces better power values for models estimated using order-2 methods (MQL2 and PQL2)

The practical application indicates the use of GOF test practically with any method of estimation. The performance of the test however varies with the selected estimation method.

V. Limitations and Suggestions

Considering the limitations of the study, one limitation is that the study considers only balanced datasets (equal number of observations within each cluster). Moreover, power of the test depends on the functional form under consideration. The simulation study of power is only based upon two such variations in standard error of the variable. Furthermore, simulations are conducted by considering only the normally distributed explanatory variable. The application of the test with other forms of explanatory variables is ignored in this study.

One suggestion for improvement of the study would be finding the applicability of the test for developed methods of estimations such as bootstrap methods and MCMC methods. Moreover, an assessment of power with varying functional forms could also be conducted and the study could be applied to unbalanced clusters. As there are many advanced GOF tests developed which takes the basis from the basic test for binary MLM, the suitability of those tests with varying methods of estimations could be evaluated.

VI. Contribution

The first author conducted simulations, did coding for simulation and the practical dataset. Moreover, the author analyzed the dataset and wrote up the paper. The second author supervised the entire study, improved codes and the paper as appropriately.

References

[1] G. Rodriguez and N. Goldman, "An assessment of estimation procedures for multilevel models with binary responses," Journal of the Royal Statistical Society, vol. 158, no. 1, pp. 73-89, 1995.

[2] J. J. Hox, Multilevel analysis- Techiniques and applications, New York: Routledge, 2010.

[3] F. L. Huang, "Alternatives to multilevel modeling for the analysis of clustered data," The Journal of Experimental Education, vol. 84, no. 1, pp. 175-196, 2016.

[4] J. Rasbash, F. Steele, W. Browne and H. Goldstein, A user's guide to MLwiN, Version 3.00, Centre for Multilevel Modelling, University of Bristol, 2017.

[5] H. Goldstein and J. Rasbash, "Improved approximations for mulitlevel models with binary respones," Journal of the royal statistical society, vol. 159, no. 3, pp. 505-513, 1996.

[6] D. Hosmer and S. Lemeshow, "Goodness of fit tests for the multiple logistic regression model," Communications in Statistics - Theory and Methods, 1980.

[7] S. Lipsitz, G. Fitzmaurice and G. Molenberghs, "Goodness-of-fit tests for ordinal response regression models," Journal of the Royal Statistical Society. Series C (Applied Statistics), vol. 45, no. 2, pp. 175-190, 1996.

[8] A. Perera, M. Sooriyarachchi and Wickramasuriya, "A goodness of fit test for the multilevel logistic model," Communications in statistics: Simulation and computation, vol. 45, no. 2, pp. 643-659, 2016.

[9] K. Archer, S. Lemeshow and D. Hosmer, "Goodness-of-fit tests for logistic regression models when data are collected using a complex sampling design," Computational statistics and data analysis, 2007.

[10] S. A. Knox and P. Chondros, "Observed intra-cluster correlation coefficients in a cluster survey sample of patient encounters in general practice in Australia," BMC Medical Research Methodology, 2004.

[11] C. Maas and J. Hox, "Sufficient sample sizes for multilevel modeling," 2005.

[12] I. G. Kreft and J. de Leeuw, Introducing multilevel modeling, Newbury Park, CA:: Sage, 1998.

[13] N. M. Huq and J. Cleland, "Bangladesh fertility survery, 1989," National Institute of Population Research and Training (NIPORT)., Dhaka, 1990.

[14] W. M. Abeysekera and R. Sooriyarachchi, "A novel method for testing goodness of fit of a proportional odds model: An application to AIDS study," Journal of National Science Foundation Sri Lanka, vol. 36, no. 2, pp. 125-135, 2008.

[15] J. Cohen, Statistical power analysis for the behavioral sciences, 2 ed., New York: Lawrence Erlbaum Associates, Publishers, 1988.

[16] J. A. Schoeneberger, "The impact of sample size and other factors when estimating multilevel logistic models," The Journal of Experimental Education, vol. 84, no. 2, pp. 373-397, 2016.