Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

H-Likelihood Estimation Method For Varying Clustered Binary Mixed Effects Model

Intesar N. El-Saeiti[(1)], Aman Pannu[(2)]

[(1)]*Department Of Statistics, Faculty Of Science, University Of Benghazi, Entesar.El-Saeiti@Uob.Edu.Ly, Benghazi-Libya*
[(2)]*Director, Analytics & Project Management, Cloudrav Inc., Aman_Pannu@Yahoo.Com*

Clustered or hierarchical data structures with binary responses are prevalent in various practical applications. These structures can involve an equal or unequal number of observations, leading to the analysis of data exhibiting intricate variability patterns. Mixed models, incorporating fixed effects of interest and random effects to address clustering, are commonly employed due to their appropriateness in practice. Random effects in these models account for multiple error structures. In the domain of clustered binary mixed effects models, the Hierarchical Generalized Linear Model (HGLM) stands out as a preferred model. This study assesses the performance of the h-Likelihood estimation method for clustered binary mixed effects models with both balanced and unbalanced cluster sizes. Evaluation through computer simulations considers parameters such as unbiasedness, Type I error rate, power, and standard error. The simulations encompass varying numbers of clusters and cluster sizes, revealing nuances in the performance of the mixed effects clustered binary data model based on the cluster sizes.

Keywords: Hierarchical Generalized Linear Model (HGLM), H-Likelihood Method, balance, unbalanced, Binary Response.

INTRODUCTION

Numerous research endeavors across health, finance, education, and the social sciences have entailed the collection of binary data organized into clusters. For instance, studies might involve the smoking status of students sampled from various schools or the disease status of animals from different farms. Such data typically exhibit correlations within clusters, where students from the same school or animals from the same farm tend to share similarities that distinguish them from individuals in other clusters. In the design of these studies, a critical decision point emerges regarding the selection of the number of groups to sample from. Opting for a larger number of groups or schools tends to reduce data dependence and enhance the precision of estimates related to explanatory variables. In certain experimental settings, clusters may be either balanced or unbalanced, with variations in the number of observations within each cluster. Unbalanced clusters can arise from uneven sub-sampling practices or random missing elements in clustered multivariate outcomes. The presence of different cluster sizes can introduce varying dispersions, thereby posing challenges of heterogeneity in models that necessitate distinct variance components, a concern previously explored in continuous response studies (El-Saeiti, 2004). This study adopts a nested design incorporating mixed effects models, a pragmatic choice due to its inclusion of fixed and random factors. When a model encompasses both fixed and random effects, it is referred to as a generalized linear mixed model (GLMM) or a hierarchical generalized linear model (HGLM) as introduced by Lee and Nelder (1996). Lee et al. (2024) provides a valuable contribution to the literature on advanced statistical methods for modeling and analyzing multivariate longitudinal binary data, leveraging the H-likelihood estimation technique. The proposed methodology can be a valuable tool for researchers and practitioners working with such complex data structures. Hierarchical generalized linear models accommodate additional error components in the linear predictors of generalized linear models, offering a non-normative distribution requirement and thereby broadening the model class. Within hierarchical generalized linear models, response variables and random effects can adhere to any distribution within the exponential family, a concept elaborated by McCullagh and Searle (2001). Consequently, HGLMs stand out as more suitable for clustered data compared to generalized linear models (GLMs). Yau, K. K., & Lee, A. H. (2021) presents a generalized mixed effects model for the analysis of longitudinal binary data. The model incorporates both random and fixed effects, and allows for the inclusion of time-varying covariates. The model parameters are estimated using the H-likelihood approach, which provides a unified framework for estimating the fixed and random effects. In generalized linear models, the estimation of the mean component typically involves using Maximum Likelihood (ML) methods. An extension to this approach within Hierarchical Generalized Linear Models (HGLM) is the Restricted Pseudo Likelihood (REPL) estimation method for binary mixed effect models, extensively discussed by El-Saeiti (2015). Comparative studies by Helena and Louise (1997) have indicated that parameter estimates obtained through ML and REPL methods exhibit fairly close agreement. To estimate both the mean and dispersion parameters, researchers often turn to the hierarchical likelihood (HL) estimation technique. Unlike traditional approaches, HL does not necessitate normality assumptions for random components, akin to the REPL method, thereby allowing for a wider array of model specifications as highlighted by Lee and Nelder (1996).

Lee and Nelder 2006, propose a class of double hierarchical generalized linear models in which random effects can be specified for both the mean and dispersion. Heteroscedasticity between clusters can be modelled by introducing random effects in the dispersion model, as is heterogeneity between clusters in the mean model.

The formulation of the hierarchical likelihood for the response variable y is expressed as:

$h = \ln(f(y|v; \beta, \varphi)) + \ln(f(v; \alpha))$.

Here, $f(y|v; \beta)$ and $f(v; \alpha)$ represent the conditional density function of y given random effect v and the density function of v, respectively. Lee and Nelder (1996) argued for developing algorithms based on the v-scale rather than the u-scale due to the flexibility of v in assuming real values, unlike u which may have constrained ranges leading to convergence issues. Parameter estimates in HGLMs are derived by maximizing the h-likelihood, leading to the computation of Maximum Hierarchical Likelihood Estimates (MHLEs). These estimates are obtained by solving the partial derivatives of the h-likelihood with respect to the fixed effects (β) and random effects (v). In the context of binary outcomes, the HGLM framework, as elucidated by Lee and Nelder (1996), involves modeling the dependent variable with a binomial distribution and the random effect with a beta distribution. Further insights into binary outcomes with beta distribution for random effects can be found in the works of El-Saeiti (2013), Lalonde (2009), and Lee and Nelder (1996).

The key components of the HGLM framework include the response distribution (Binomial), random distribution (Beta), linear component (η), and the link function (logit). The h-likelihood for the binomial-beta model is given by:

$h = l(\beta, \varphi; y|v) + l(\alpha; v).$

The estimation equations for the fixed part (β) and random component (v) in the h-likelihood estimation process are derived to obtain estimates for both parameters, ensuring a comprehensive understanding of the model and its components.

SIMULATION

In the simulation study conducted, the researcher initiated data generation by creating two distinct datasets: one with balanced cluster sizes and the other with unbalanced cluster sizes. Parameters were defined and values were generated including random effect variables, followed by the calculation of probabilities for the dependent variable. In cases of unequal cluster sizes, varying numbers of subjects were generated per cluster using a Poisson distribution, where the mean for the Poisson distribution represented the average number of observations within each cluster. By altering the mean cluster sizes ($\bar{n}$ = 10, 25, 50, 100), the researcher illustrated the impact on statistical performance across different sample sizes.
Furthermore, a normally distributed continuous variable, xi j, was generated with a mean of 3 and a known variance of 20 ($x_{ij} \sim N(3, 20)$). Subsequently, a beta-distributed random variable, ui, was created with parameters $\gamma = 2$ and $\lambda = 3$ for each cluster i ($u_i \sim$ Beta(2, 3)). For scenarios with equal cluster sizes, similar processes were followed, but with an equal number of observations in each cluster.
Each data unit was randomly generated from a Bernoulli distribution with a success probability calculated as $p_{ij} = \frac{e^{(\beta_0 + \beta_1 x_{ij} + u_i)}}{1+ e^{(\beta_0 + \beta_1 x_{ij} + u_i)}}$.
Here, $\beta_0 = 1$ and $\beta_1 = 0.2$, and parameter estimates were derived using the H-Likelihood method Heo and Leon (2005).

The study specified the number of clusters (K = 10, 20, 50, 100), the cluster size for balanced clusters (n = 10, 25, 100), and for unbalanced clusters, the mean number of observations per cluster ($\bar{n}$ = 10, 25, 100). For each combination of K and n, 1,000 datasets were generated for both equal and unequal cases to evaluate power, Type I error rates, and standard errors. Power, Type I error rates, and standard errors were computed based on the model with the systematic component $\eta_{ij} = \beta_0 + \beta_1 x_{1ij} + v_i$, with a specified treatment effect for β_1.
The H-Likelihood Hierarchical Generalized Linear Model (HGLM) was utilized for data generation, where the systematic component for data generation was $\eta_{ij} = 1 + 0.2\ x_{1i\,j} + v_i$, and for the model fitting, it was represented as $\eta_{ij} = 1 + 0.2\ x_{1ij} + 3.1\ x_{2ij} + vi$, with vi $\sim$ Beta(2, 3). The researcher employed the HGLM function within the HGLM package in R to estimate parameters β and t-statistics along with corresponding p-values. By averaging 1,000 estimates

obtained through simulation, the researcher calculated the values for β₁, β₂, power of the hypothesis test for β₁, Type I error rate for β₂, and the standard error for β₁.

RESULT:

Table 1 for Binomial Beta h-likelihood estimate parameters. The Binomial Beta h-likelihood estimate

		Balanced Cluster		Unbalanced Cluster	
Clusters	Sample size	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$
K = 10	10	0.2319765	-0.007228321	0.1958833	0.009286461
	25	0.1939059	0.003553967	0.2017746	0.0108503
	50	0.1970002	-0.002042296	0.188225	-0.0001238602
	100	0.199145	0.002284678	0.2009817	-0.01050844
K = 20	10	0.215392	-0.03054897	0.210038	0.01873527
	25	0.2038395	-0.01017131	0.2013315	-0.001884942
	50	0.2035105	0.004907986	0.2022876	0.0006811804
	100	0.2006388	-0.002680622	0.1983477	-0.000997808
K = 50	10	0.2080814	0.001532905	0.1958833	0.009286461
	25	0.1994717	0.002696468	0.2022252	0.006061514
	50	0.1967751	-0.0005004571	0.2000865	0.002234016
	100	0.2001256	0.0007905866	0.20241	0.000397104
K = 100	10	0.2004939	0.001584383	0.196161	0.003048525
	25	0.2016236	-0.002657747	0.202098	0.002534502
	50	0.1991661	0.0008547018	0.2014994	0.001459892
	100	0.1996344	-0.00128299	0.1980433	0.001697924

Table 1: Estimate parameters

The estimation of parameters β1 and β₂ using Binomial Beta h-likelihood for both balanced and unbalanced cluster sizes demonstrated values that closely approximated the true parameters, with β1 estimated at 0.2 and β2 at 0. The Binomial Beta h-likelihood method proved to be effective in providing estimates that closely matched the actual values.

In Table 2, the Binomial Beta h-likelihood Type I error rates for β₂ were detailed for both balanced and unbalanced cluster sizes. Type I error rates were calculated as the proportion of p-values less than 0.05 under the null hypothesis H₀: β₂ = 0. Ideally, the Type I error rate should hover around 0.05. The explanation of the Type I error rate for β₂ revealed slightly varying values for equal and unequal cluster sizes. It was observed that balanced cluster sizes exhibited lower values compared to unbalanced cluster sizes.

Table2: Type I Error

Clusters	Sample size	Balanced	Unbalanced
K = 10	10	0.12	0.085
	25	0.07	0.095
	50	0.12	0.09
	100	0.073	0.104
K = 20	10	0.136	0.109
	25	0.09	0.096
	50	0.165	0.108
	100	0.067	0.087
K =50	10	0.067	0.085
	25	0.065	0.126
	50	0.087	0.104
	100	0.123	0.089
K = 100	10	0.102	0.06
	25	0.082	0.134
	50	0.095	0.136
	100	0.087	0.121

Table 3 illustrated the power of the hypothesis test for β_1 using the Binomial Beta h-likelihood method. Statistical power was determined as the ratio of correctly rejected null hypotheses (H_0: $\beta_1 = 0$). The test was iterated 1,000 times through simulation to ascertain how frequently the test yielded significant results. Power represented the proportion of these 1,000 tests that were correctly rejected.

It was observed that balanced cluster sizes exhibited greater statistical power compared to unbalanced cluster sizes, particularly evident with smaller sample sizes. The power statistics for balanced clusters surpassed those for unbalanced clusters, indicating that the Binomial Beta h-likelihood method provides more accurate estimates for balanced cluster binary models than for unbalanced cluster models.

Table 3: Power

Clusters	Sample size	Balanced	Unbalanced
K =10	10	0.89	0.906
	25	1	0.677
	50	1	0.864
	100	1	0.991
K =20	10	0.998	0.615
	25	1	0.937
	50	1	0.999

	100	1	1
K =50	10	1	0.906
	25	1	1
	50	1	1
	100	1	1
K =100	10	1	0.991
	25	1	1
	50	1	1
	100	1	1

In Table 4, the concept of Standard Error (SE) is examined. The average Standard Error ($\overline{\text{SE}}$) was determined as the mean of the 1,000 SE values for the estimates of β_1. A smaller $\overline{\text{SE}}$ denoted reduced estimated variability or increased precision in the parameter estimates. The standard error for $\hat{\beta}$ indicated the level of efficiency improvement.

The findings in Table 4 suggest that the Binomial Beta h-likelihood method exhibited smaller standard errors for balanced clusters.

Table 4: Stranded error

Clusters	Sample size	Balanced	Unbalanced
K =10	10	0.07152838	0.05695932
	25	0.04197166	0.08128032
	50	0.02903917	0.05683201
	100	0.0202115	0.04005908
K =20	10	0.04737441	0.09272815
	25	0.02885089	0.05658015
	50	0.02028826	0.04003575
	100	0.0142676	0.02824783
K =50	10	0.02903625	0.05695932
	25	0.01807183	0.03579394
	50	0.0127137	0.02526456
	100	0.00901145	0.01782909
K =100	10	0.0202617	0.04016537
	25	0.01277624	0.0252753
	50	0.009003529	0.01786361
	100	0.006371349	0.01261467

Tables 1 through 4 present a comprehensive overview of the simulation results for the Binomial Beta h-likelihood method applied to both equal and unequal cluster sizes. These tables summarize the findings related to parameter estimation, power statistics, Type I error rates, and standard errors. The analysis from these tables indicates that the Binomial Beta h-likelihood method serves as a reliable estimator. Across 1,000 replications, the estimates were remarkably close to the true values, with β_1 estimated at 0.2 and β_2 approximating zero. In terms of statistical power, balanced clusters exhibited higher values compared to unbalanced clusters, while Type I error rates were notably lower for balanced clusters than for unbalanced ones.

Moreover, smaller average standard errors (SE) in the estimation process indicated reduced variability and enhanced precision in parameter estimates. Notably, balanced cluster sizes displayed superior performance compared to unbalanced cluster sizes in terms of these metrics, highlighting the efficacy of the Binomial Beta h-likelihood method, particularly for balanced cluster binary models.

CONCLUSIONS:

The Binomial Beta h-likelihood method emerged as a robust approach for handling mixed effects in clustered binary data models, showcasing nuanced variations based on cluster sizes. Through 1,000 replications, the estimates closely mirrored the actual values, demonstrating the method's effectiveness. Notably, in balanced scenarios, the power of the hypothesis test for regression parameters outperformed unbalanced setups, while the Type I error rates for these tests were deemed acceptable, notably lower for balanced clusters compared to unbalanced ones. Furthermore, the standard error associated with regression parameters was minimal.

This study establishes the Binomial Beta h-likelihood method as a viable estimation technique, particularly well-suited for balanced clustered sizes over unbalanced cluster binary responses. The simulation results underscore the method's proficiency, especially in scenarios with balanced cluster sizes.

FUTURE WORK

Since Binomial Beta h-likelihood is an acceptable estimation method for balanced clustered sizes more than unbalanced clusters binary response; It is a good idea to adjust the Binomial Beta h-likelihood estimate method to deal with unbalanced cluster size which will be the next work for the author.

References

a. El-Saeiti, I. N. (2004). Messy data in heteroscedastic models case study: Mixed nested design. M.Sc.THESIS. university of Benghazi.

b. El-Saeiti, I. N. (2013). Adjusted variance components for unbalanced clustered binary data models. Ph.D. Dissertations. University of Northern Colorado.

c. El-Saeiti, I. N. (2015). Performance of mixed effects for clustered binary data models. AIP Conf. Proc.,1643:, 80–85.

d. Gu, Z. (2008). Model diagnostics for generalized linear mixed models. Dissertations.

e. Helena, Geys. Geert, M. and Louise, M. R. (1997). Pseudo-likelihood inference for clustered binary data. COMMUN STATIST-THEORY METH, 26(11):2743–2767.

f. 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.

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

h. Lee, K.-J., Kim, C., Yoo, J. K. and Lee, K.* (2024). Multivariate probit linear mixed models for multivariate longitudinal binary data. Statistics in Medicine.

i. Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. Journal of the Royal Statistical Society, Series B (Methodological), 58(4):619–678.

j. Lee, Y. and Nelder, J. A. (2006). Journal of the Royal Statistical Society Series C: Applied Statistics, Volume 55, Issue 2, April 2006, Pages 139–185, https://doi.org/10.1111/j.1467-9876.2006.00538.x

k. McCullagh, C. E. and Searle, S. R. (2001). Generalized, Linear, and Mixed Models. John Wiley & Sons,Inc., New York.

l. Yau, K. K., & Lee, A. H. (2021). A generalized mixed effects model for longitudinal binary data. Statistical Methods in Medical Research, 30(1), 82-97.