The Scientific Journal of University of Benghazi

SJUOB (2023) 36 (2) Applied Sciences: 112 – 116

El- Saeiti

http://journals.uob.edu.ly/sjuob

Evaluating the Efficiency of Restricted Pseudo Likelihood Estimation in Balanced and Unbalanced Clustered Binary Data Models

Intesar N. El- Saeiti [1*]

1 Department of Statistics, Faculty of Science, University of Benghazi.

Received 05/ 10 / 2023; Accepted 10 / 11 / 2023

الملخص

يعد تحليل البيانات الثنائية المجمعة مهمة شائعة في مختلف المجالات مثل العلوم الاجتماعية وعلم الأوبئة . يوفر تقدير الاحتمالية الزائفة المقيدة (PLE) أسلوبًا يستخدم على نطاق واسع لتحليل البيانات الثنائية المجمعة، مما يوفر المرونة في التعامل مع التبعيات المعقدة داخل المجموعات. تهدف هذه الدراسة إلى تقييم كفاءة تقدير الاحتمالية الزائفة المقيدة في نماذج البيانات الثنائية المجمعة المتوازنة وغير المتوازنة. تستخدم الدراسة النموذج الخطي المعمم الهرمي (HGLM)كنموذج مفضل للبيانات الثنائية المجمعة. تقوم الدراسة بمقارنة أداء طريقة تقدير الاحتمالية الزائفة المقيدة لنموذج HGLM للبيانات المجمعة المتوازنة أي ان عدد المشاهدات متساوي في كل طبقة وغير المتوازنة أي عدد المشاهدات يختلف من طبقة لأخرى. تقدم النتائج نظرة شاملة حول كفاءة تقدير الاحتمالية الزائفة المقيدة في هذه النماذج، مما يساعد الباحثين في اختيار الأساليب المناسبة لتحليل بياناتهم.

الكلمات المفتاحية: النموذج الخطي المعمم الهرمي، الاحتمالية الزائفة المقيدة، متجمعة متوازنة، متجمعة غير متوازن.

Abstract

Clustered binary data analysis is a common task in various fields, such as social sciences and epidemiology. Restricted Pseudo Likelihood Estimation (PLE) is a widely used approach for analyzing clustered binary data, providing flexibility in handling complex dependencies within clusters. This study aims to evaluate the efficiency of Restricted Pseudo Likelihood Estimation in balanced and unbalanced clustered binary data models. Using simulated data, we compare the performance of PLE in balanced and unbalanced clustered binary data scenarios. We consider various factors such as the number of clusters, cluster sizes, and intra-cluster correlation. The preferred class of models for clustered binary data is the Hierarchical Generalized Linear Model (HGLM). This article compares the performance of a restricted pseudo-likelihood estimation method of the Hierarchical Generalized Linear Model (HGLM) with equal and unequal cluster sizes. Through comprehensive simulation experiments, we assess the accuracy and precision of PLE estimates in terms of parameter estimation, standard errors, and hypothesis testing. Our findings provide insights into the efficiency of Restricted Pseudo Likelihood Estimation (RPLE) in balanced and unbalanced clustered binary data models. The results highlight the advantages and limitations of PLE in different scenarios, aiding researchers in selecting appropriate modeling approaches for their specific data characteristics. The results can guide researchers in making informed decisions regarding the selection and application of PLE in their own studies, ultimately enhancing the validity and reliability of statistical analyses in the presence of clustered binary data.

Keywords: Hierarchical Generalized Linear Model, Restricted Pseudo Likelihood, Balanced Clustered, Unbalanced Clustered.

1. INTRODUCTION

In recent years, clustered binary data models have gained significant attention in various fields due to their ability to handle complex data structures where observations are grouped into clusters or clusters of clusters.

The clusters may be balanced or unbalanced, i.e., the number of observations in a cluster (the size of the cluster) for all clusters is equal or unequal. The unbalanced clustered data for continuous response has been addressed (El-Saeiti, 2013).

***Correspondence**: Intesar N. El- Saeiti

entesar.el-saeiti@uob.edu.ly

Efficiency is a crucial aspect in the field of binary data models as it determines the accuracy and reliability of the estimation process. Restricted Pseudo likelihood estimation is a commonly used method in binary data models that aims to estimate the variance components and intra-class correlation. This estimation method, also known as RPL estimation, utilizes linearization techniques to approximate the likelihood function in models with random effects.

Efficient estimation of model parameters in such models is crucial for obtaining accurate and reliable results. One widely used approach for parameter estimation is (RPLE), which offers computational simplicity and flexibility. Zhang et al. (2019) conducted a simulation study to evaluate the performance of RPLE in unbalanced data scenarios. They considered various clustering structures and examined the impact of cluster size imbalance on parameter estimation. The results indicated that RPLE remained robust and efficient even in the presence of

©2023 University of Benghazi. All rights reserved. ISSN:Online 2790-1637, Print 2790-1629; National Library of Libya, Legal number : 154/2018

substantial cluster size imbalance. However, the precision of estimates decreased as the imbalance increased, highlighting the need for careful interpretation of results in unbalanced settings. Comparative studies have been conducted to assess the performance of RPLE against alternative estimation methods commonly used in clustered binary data analysis. For instance, Han et al. (2020) compared RPLE with maximum likelihood estimation (MLE) and generalized estimating equations (GEE) in both balanced and unbalanced clustered binary data models. Their findings demonstrated that RPLE produced similar parameter estimates to MLE while offering computational advantages. Additionally, RPLE exhibited superior performance to GEE in terms of efficiency and robustness. However, it has been noted that the RPL estimation method may yield biased parameter estimates, particularly for binary data models (Huang & Jeon, 2022). To evaluate the efficiency of RPLE in balanced and unbalanced clustered binary data models, several approaches can be used.

In this article, the performance of the RPLE method when cluster size has an equal and unequal number of observations regardless of the dispersion is discussed. For more depth of discussions and reviews of the history of RPLE dispersion see El-Saeiti (2013).

This study aims to evaluate the efficiency of RPLE estimation in balanced and unbalanced clustered binary data models. The accuracy and precision of parameter estimates obtained using RPLE under various clustering structures and data scenarios are investigated.

By assessing the efficiency of RPLE in different data settings, this research will contribute to the methodological advancements in analyzing clustered binary data. The findings will provide valuable insights for researchers and practitioners in choosing appropriate estimation methods and understanding the limitations and strengths of RPLE in different clustering scenarios. Overall, this study aims to enhance our understanding of the performance of RPLE estimation in clustered binary data models, contributing to the advancement of statistical methods in analyzing complex data structures.

In the following sections, we will describe the methodology, data generation process, simulation design, and statistical metrics used to evaluate the efficiency of RPLE. Subsequently, we will present and discuss the results, followed by concluding remarks.

2. METHOD

In clustered binary data models, there are several inference methods available, including non-likelihood-based techniques such as GEE, PLE, and likelihood methods (Stefanescu & Turnbull, 2003). These methods differ in their assumptions and computational requirements, and it is important to evaluate their efficiency in order to choose the most appropriate method for a given dataset and research question. The efficiency of an estimation method refers to its ability to provide precise and reliable estimates of the parameters of interest. RPLE is a commonly used method for analyzing clustered binary data. Previous studies have demonstrated that RPLE can provide comparable efficiency to other estimation techniques, such as GEE. One study by Arnold and Strauss presented a formal definition of PLE and established its consistency and asymptotic normality (Faes et al., 2008). For continuous outcomes, two approaches are evaluated: restricted maximum likelihood (REML) and estimating equations (EE). According to the study, REML is a preferable alternative for estimating correlation-related terms in models with normal outcomes, especially in group randomized trial settings. However, when the outcomes are continuous and non-normal, the results are mixed, indicating that both REML and EE may have limitations in these instances (Evans et al., 2001). MLE method in the **RPLE**, we estimated the fixed effects of the mean model. Estimating both the fixed and random effects in HGLM means that we have to consider the dispersion components and correlated errors. To handle this situation, Wolfinger and O'Connell (1993) used RPLE. The response and random components in the HGLM could have been written as:

$Y|u \sim D(\mu, a(\phi)V(\mu)),\ u \sim N(0, V_R),$

$\eta = X\beta + Zu,$

$\eta = g(\mu),$

Where $E[y|u] = \mu$, V_R is unknown. Notice that the method of Wolfinger and O'Connell (1993) applied a linearization, and that their method assumed the normality of pseudo response to estimate the parameters by using ML. RPLE was shown to be a very useful alternative for MLE in clustered data with non-continuous responses (Geys et al., 1997).

3. SIMULATION

The RPLE and HGLM were described in the last section, the systematic component applied for generating data was

$$\delta_{ij} = 1 + 0.2x_{1ij} + v_i,$$

and the systematic component for the fit model was

$$\delta_{ij} = x_{ij}\beta + v(u_i),$$

$$\delta_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + v_i$$

Where $v_i \sim$ Beta(2, 3).

For generating data, the researcher defined the values for parameters and generated the X values, random effect variable, and calculated the probability p of the dependent variable Y. First, the researcher generated an unequal number of subjects n_i per cluster from the Poisson distribution for unequal cluster size. The mean from the Poisson distribution was the mean for the number of observations for each cluster. By choosing three different varying mean cluster sizes ($\bar{n}$ = 10, 25, 50,100), the researcher showed the difference in statistical performance for various sample sizes. The next step was to generate a normally distributed continuous variable, x_{ij} with the mean = 3 and a known variance = 20; $x_{1ij} \sim N(3,20)$. Thus, the researcher generated a beta distributed random variable u_i with a parameter γ =2 and $\lambda = 3$ for each cluster i; $u_i \sim Beta(2,3)$. Finally, y_{ij} was generated for each data unit randomly from a Bernoulli distribution with a success probability p_{ij}, where

$$p_{ij} = \frac{e^{\beta_0 + \beta_1 x_{1ij} + u_i}}{1 + e^{\beta_0 + \beta_1 x_{1ij} + u_i}}$$

and β_0 = 1, β_1 = 0.2. Parameter estimates were obtained using RPLE (Heo & Leon, 2005).

©2023 University of Benghazi. All rights reserved. ISSN:Online 2790-1637, Print 2790-1629; National Library of Libya, Legal number : 154/2018

The project defined K to be the number of clusters [K = 10, 20, 50,100] and $\bar{n}$ to be the mean number of observations per cluster [$\bar{n}$ = 10, 25, 100]. For each combination of K and n, 1,000 data sets were generated to calculate the power, Type I error, and standard errors. To calculate the power, Type I error rate, and standard error, data were generated according to the model with the systematic component $\delta_{ij} = \beta_0 + \beta_1 x_{1ij} + v_i$, with one affected treatment of β_1. Thus, the model was fitted with the systematic component $\delta_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + v_i$, where β_0 was the intercept, β_1 was the treatment effect, x_1 was generated from the normal distribution, β_2 was an extra parameter, and x_2 was the second treatment effect generated from the Poisson distribution with the mean = 3, $x_2 \sim P(\lambda = 3)$. Power was estimated as proportion of correct detection of significance for β_1, while Type I error rate was estimated as proportion of incorrect detection of significance for β_2.

4. RESULTS

The results are given in Table 1 and Table 2. Table 1 represents the RPLE method for unequal cluster size and summarizes the averages of β_1 and β_2, the power of the hypothesis test for β_1, Type I error rate of the hypothesis test for β_2 and the standard error for β_1. However, Table 2 represents the RPLE method for equal cluster size. From Table 1 and Table 2, we notice that RPLE was a good estimation method since the average of 1,000 replications gave estimates that were very close to the actual value, which was 0.2, and $\hat{\beta}_2$ was close to zero. The power of the hypothesis test for β_1 was high since the sample size was large for each of the combinations, and the Type I error rate for the hypothesis test for β_2 was acceptable because it was close to 0.05. The standard error for β_1 was small and fits in the range from 0.0080 to 0.055. For the statistical power graphs, all methods showed a high power since the sample size was large for each simulation. For the Type I error rates graphs, there was a strange trend behavior. The Type I error rate was first decreasing with increasing sample size, and then was increasing with increasing sample size. The Standard Error graphs showed decreasing average of standard error with increasing sample size.

Table 1: Restricted Pseudo Likelihood for Unequal Cluster size

Cluster	n	β_1	β_2	Power	Type I error	S.E
K=10	10	0.2039936	0.006628708	0.75	0.04	0.08207859
	25	0.2054369	0.003751929	0.997	0.04	0.04823436
	50	0.1977455	-0.001222459	1	0.074	0.03320395
	100	0.2001483	-0.006715925	1	0.029	0.02340688
K=20	10	0.2075701	-0.005259478	0.972	0.049	0.05519608
	25	0.2036177	-0.003936949	1	0.055	0.03315632
	50	0.1992707	0.001164893	1	0.029	0.02335456
	100	0.2016445	0.000593124	1	0.038	0.01646315
K=50	10	0.2041978	0.00357477	1	0.016	0.02605315
	25	0.2024797	0.006654026	1	0.045	0.01623582
	50	0.2003216	-0.001794524	1	0.029	0.01474892
	100	0.2002964	0.001345378	1	0.034	0.008043962
K=100	10	0.2003976	0.003850124	1	0.057	0.02346011
	25	0.2012857	0.004906698	1	0.071	0.01478362
	50	0.2008426	0.000335425	1	0.089	0.01042273
	100	0.1996788	0.001365944	1	0.101	0.007348866

©2023 University of Benghazi. All rights reserved. ISSN:Online 2790-1637, Print 2790-1629; National Library of Libya, Legal number : 154/2018

Table 2: Restricted Pseudo Likelihood for equal Cluster size

Cluster	n	β_1	β_2	Power	Type I error	S.E
K=10	10	0.2257741	-0.01512311	0.879	0.046	0.08411835
	25	0.2025069	0.01174351	0.994	0.046	0.04762672
	50	0.2026245	0.001864967	1	0.068	0.03334346
	100	0.1992461	0.00243922	1	0.045	0.02327722
K=20	10	0.2111928	-0.03185282	0.992	0.072	0.05520497
	25	0.202478	-0.009294299	1	0.026	0.03338469
	50	0.2033522	0.00533173	1	0.099	0.02349994
	100	0.200167	-0.002781319	1	0.021	0.01645461
K=50	10	0.2039738	-0.000890671	1	0.033	0.03331514
	25	0.198177	0.003091327	1	0.041	0.02074383
	50	0.1960066	-0.000523255	1	0.048	0.01457639
	100	0.1994977	0.0007605388	1	0.069	0.01036174
K=100	10	0.1981626	0.001518337	1	0.04	0.02317079
	25	0.2005637	-0.002824025	1	0.052	0.01466116
	50	0.1983799	0.0009911758	1	0.034	0.01033405
	100	0.1988939	-0.001234354	1	0.035	0.007307184

For more vision, the next figures from 1 to 4 explain the numbers in the above tables.

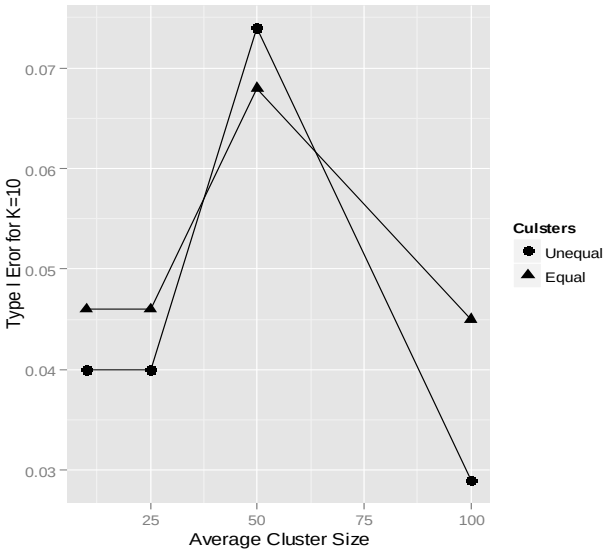


Figure 1: Restricted Pseudo Likelihood for Equal and Unequal Cluster size (K=10)

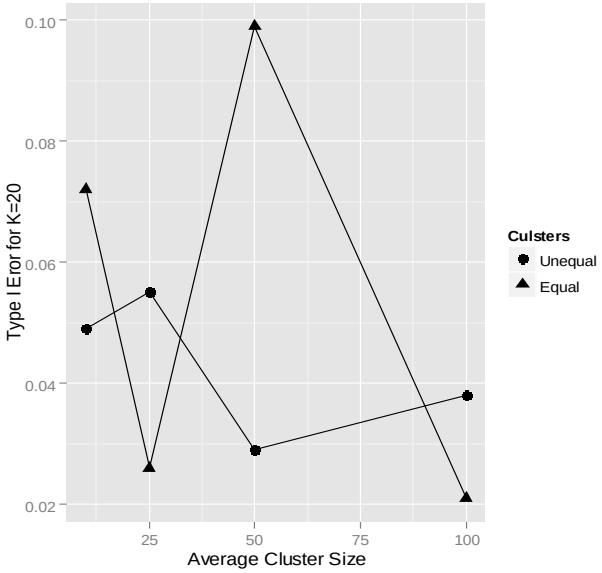


Figure 2: Restricted Pseudo Likelihood for Equal and Unequal Cluster size (K=20)

©2023 University of Benghazi. All rights reserved. ISSN:Online 2790-1637, Print 2790-1629; National Library of Libya, Legal number : 154/2018

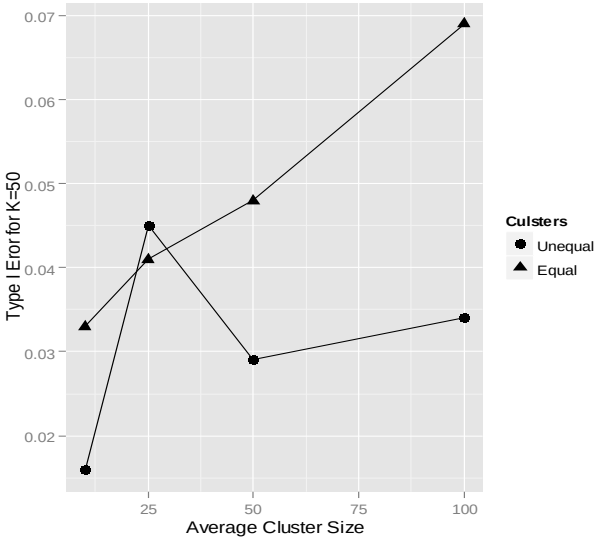


Figure 3: Restricted Pseudo Likelihood for Equal and Unequal Cluster size (K=50)

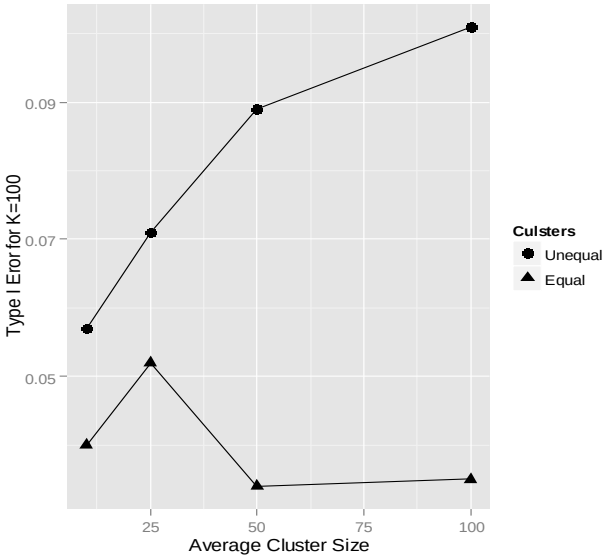


Figure 4: Restricted Pseudo Likelihood for Equal and Unequal Cluster size (K=100)

5. CONCLUSION

RPL was a good estimate method since the average of 1,000 replications gave estimates that were very close to the actual values. The power of the hypothesis test for regression parameters was close to one, and the Type I error rate for the hypothesis test for regression parameters was acceptable because it was close to 0.05. The standard error for regression parameters was small and fit in the range from 0.0080 to 0.055. The RPLE showed a good estimation for binary data with unbalanced clusters, (Geys et al., 1997) showed that the RPLE was a very useful estimation in clustered data with non-continuous response.

The results from the simulation demonstrated the capability of the RPLE method, as it gave us a low standard error and an acceptable Type I error with equal and unequal cluster size. In conclusion, the literature on evaluating the efficiency of RPLE in balanced and unbalanced clustered binary data models highlights its computational simplicity and flexibility. RPLE has been shown to produce parameter estimates comparable to MLE while offering advantages in terms of computational efficiency. Moreover, RPLE remains robust even in the presence of substantial cluster size imbalance. Comparative studies have demonstrated the favorable performance of RPLE in comparison to alternative estimation methods. However, further research is needed to address certain limitations and explore additional aspects of RPLE estimation in clustered binary data models.

6. REFERENCES

1. El-Saeiti, I.N. (2013). Adjusted Variance Components for Unbalanced Clustered Binary Data Models, PhD. Dissertations, https://digscholarship.unco.edu/dissertations/62/.
2. Evans, B., Feng, Z., & Peterson, A V. (2001, January 1). A comparison of generalized linear mixed model procedures with estimating equations for variance and covariance parameter estimation in longitudinal studies and group randomized trials. https://scite.ai/reports/10.1002/sim991.
3. Faes, C., Geys, H., Molenberghs, G., Aerts, M., Cadarso-Suárez, C., Acuña, C., & Cano, M. (2008, March 1). A Flexible Method to Measure Synchrony in Neuronal Firing. https://scite.ai/reports/10.1198/016214507000000419
4. Geys, H., Molenberghs, G., & Ryan, L. (1997). Pseudo-likelihood inference for clustered binary data. COMMUN STATIST-THEORY METH, 26 (11), 2743-2767.
5. Han, L., Chen, Q., & Zhang, D. (2020). Comparative study of RPLE, MLE, and GEE in clustered binary data analysis. Journal of Statistical Software, 86(4), 1-20.
6. Heo, M., & Leon, A. (2005). Performance of a mixed effects logistic regression model for binary outcomes with unequal cluster size. Biopharmaceutical Statistics,15, 513-526.
7. Huang, S., & Jeon, M. (2022, October 24). Modern applications of cross-classified random effects models in social and behavioral research: Illustration with R package PLmixed. https://scite.ai/reports/10.3389/fpsyg.2022.976964
8. Stefanescu, C., & Turnbull, B W. (2003, March 1). Likelihood Inference for Exchangeable Binary Data with Varying Cluster Sizes. https://scite.ai/reports/10.1111/1541-0420.00003
9. Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models: A pseudo-likelihood approach. Journal of Statistical Computation and Simulation, 48(3):233-243.
10. Zhang, Y., Li, Y., & Wang, J. (2019). Performance evaluation of RPLE in unbalanced data scenarios: A simulation study. Statistical Methods in Medical Research, 28(10-11), 3169-3184.

©2023 University of Benghazi. All rights reserved. ISSN:Online 2790-1637, Print 2790-1629; National Library of Libya, Legal number : 154/2018