An adjusted scale binomial Beta H-Likelihood estimation method for unbalanced clustered binary response models

Intesar N. El-Saeiti

Department of Statistics, Faculty of Science, University of Benghazi, Benghazi-Libya

E-mail address: entesar.el-saeiti@uob.edu.ly

Highlights

- Comparing an Adjusted Scale Binomial Beta H-likelihood method and Binomial Beta H-likelihood method for dealing with over-dispersion for unequal cluster binary data models.
- By using simulation technique the adjusted method gave a slightly different results compared to the original ”existing” Binomial Beta H-likelihood method.

ABSTRACT

In practice, clustered binary responses are very prevalent, where binary data is naturally grouped by sampling techniques. Clusters are often unequal in size in some areas of studies, such as medicine, education and others. The most suitable models for binary data clusters of unbalanced sizes are the Hierarchical Generalized Linear Model (HGLM), where the random term over-dispersion counts; and it is k known as clustered binary data. Current techniques for estimating parameters in (HGLM) are many, but these techniques do not allow over dispersion to be distinct from cluster to cluster. When clustered binary data resulted in over-variation, that reasonable to conclude the unequal size of clustered binary data may have been distinct variations for distinct clusters. By ignoring the chance of shifting over variability between clusters, test statistics may be inflated in the Type I error rates. In this paper, the binomial beta (BB) (HGLM) method has been altered to account for distinct variations across separate clusters. In order to explore whether the Adjusted Scale Binomial Beta (ASBB) method is more suitable than the (BB) technique for dealing with over-dispersion for unequal cluster binary data models, the author used simulation, the adjusted method was compared to the original “existing” technique in terms of, Type I error rate, estimator standard errors and power. (ASBB) h-likelihood “adjusted” method was comparable to BB "existing" technique, as it has a less standard error and the Type I error was acceptable. Moreover, Type I error inflated in “exist method” (BB) h-likelihood.

1. Introduction

The nested structure or the clustered data is one of the experimental designs where the variables have an implicit hierarchy. Clusters may be balanced or unbalanced, which means that the size of the cluster is equal or unequal. There are many explanations for unequal clusters; see for more information on unbalanced (Milliken et al., 1992). The differential size of clusters may lead from randomly missing vector components for a clustered multivariate outcome or if fields vary in the number components in the vector for evaluation. Different cluster sizes can lead to varying cluster dispersions. There may be two sources of variation for a nested model with a binary response. The first source of variation is the variation between clusters, which represents the variation from cluster to cluster. The second is the intra-cluster variation that reflects the random variation between the responses in each cluster. For binary data that are clustered with variation at each stage, the linear model used, which assumes that the dependent variable is normal, it would be more appropriate to use the linear model extension, which is a generalized linear model. The generalized linear model (GLM) includes dependent variables that follow any probability distribution in the exponential family of distributions. The exponential family has many useful distributions for example Normal, Binomial, Poisson, Multinomial, Gamma, Negative Binomial, and others, for more details see (McCullagh and Searle, 2001). Assuming a normal distribution is convenient, but it is not always the best choice in a HGLM (Lee and Nelder, 1996). Applied hypothesis tests in the GLM do not require normality for dependent variable, nor do they require homogeneity of variances. Hence, GLMs can be used when the dependent variables follow any distributions other than the normal distribution and the variances are not constant; more details in (El-Saeiti, 2013).

Cluster design with binary outcomes is very common in study fields, particularly in medical studies. The nested structure with an unbalanced cluster size may lead to more variability between clusters. The hierarchical generalized linear model (HGLM) technique is used to account for additional variability caused by distinct cluster sizes. The most popular techniques, such as quasi-likelihood, penalized quasi-likelihood, and extended quasi-likelihood, allow for over-dispersion; however, present techniques handle over-variation as a constant for all clusters. It is common not to apply these methods to modifications in over-dispersion. Unqualified clustered binary data may have separate cluster dispersions. It is prevalent not to apply these techniques to changes in over-dispersion. Unqualified clustered binary data may have separate dispersions for separate clusters. It is sensible to believe that unequal clustered binary data may have distinct dispersion for distinct clusters, but the present techniques have ignored this option. El-Saeiti, (2014) and El-Saeiti, (2015) proved the current HGLM methods do not deal with different dispersion for different clusters. By neglecting to account, for different dispersion in binary data
with unequal clusters, Type I error rate may be inflated, power may be low and efficiency may be lost. The author modified Binomial Beta h-likelihood method for solving the problem. Adjusted Scale Binomial Beta (ASBB) to account for over-dispersion in unequal clustered binary data better than current Binomial Beta (BB) h-likelihood techniques. The adjusted Scale Binomial Beta h-likelihood enables a distinct scale parameter for the Beta distribution for each cluster to account for over-dispersion.

2. Theoretical Background

In generalized linear models (GLM) where the model includes both fixed and random effects, it is referred to as generalized linear mixed models (GLMM) or hierarchical generalized linear models (HGLM) (Lee and Nelder, 1996). Hierarchical generalized linear models enable additional error parts in the linear predictors of generalized linear models. The distribution of these components is not needed to be normal, enabling a wider class of models. In generalized hierarchical linear models, any distribution in the exponential family may be followed by response and random effects. As such, the HGLM is more suited to clustered data than the GLM.

By assuming that the conditional dependent variable $Y_{ij}$ is binomial, and assuming that beta distribution for the random effect, the distribution of the conditional response and the random effect is fully defined; in this case, the appropriate estimation method is $h$-Likelihood (Lee and Nelder, 1996).

The HGLM formula for Binomial Beta $h$-likelihood according to (Lee and Nelder, 1996) is

1. \( Y_{ij} \sim Bin(n_i, p_i) \), \( i = 1, \ldots, K \)
2. \( \eta_{ij} = x_{ij} \beta + v(u_i) \)
3. \( \eta_{ij} = \text{logit}(p_i) \)

Where $Y$ is dependent variable follow binomial distribution with parameters $n$ and variance-covariance $p$. The parameter $u$ is the random effect following the beta distribution with mean equal to $\lambda_i$ and $\lambda_i$ is the varying scale from cluster to cluster. $\eta_{ij}$ is the systematic component, and $v$ is the transformation of $u_i$ to occur linearly with $x_{ij} \beta$. $\beta$ is the fixed-parameter, $x_{ij}$ is explanatory variable for fixed effects $j$th observation in $i$th cluster, and $g$ is the link function which is logit for binomial distribution; more details are in the dissertations of El-Saeti, (2013) and Lalonde, (2009). The objective of this modified method is to allow dispersion to differ in clusters of distinct sizes and to allow variations to differ from cluster to cluster instead of a steady amount of variations, which is one.

3. Material and Methods

The author generates two data sets, one for the "original method" Binomial Beta BB method, and the second data set for the "modified method" Adjusted Binomial Beta; defined parameters and generated values, random effect variable, and calculated the probability of the response variable. The distribution of Poisson generated an unequal number of topics per cluster for the unbalanced size of the cluster. Where the mean for the Poisson distribution was the mean for the number of observations for each cluster. By selecting separate mean cluster sizes ($i=10, 25, 100$), the writer reveals the distinction in statistical output for the distinct sample sizes. In this document, the described number of clusters ($K=20,50$), the cluster size for the unbalanced cluster is the mean number of cluster observations per cluster. For each combination of cluster number $K$ and observation numbers "n", 1,000 data sets were generated for each case (BB) and (ABB) for the calculation of power, Type I error and standard error. Revise El-Saeti, (2013) for explanation of the simulation steps. Power was estimated as the percentage of correct significance detection for $\beta_1$, while the rate for Type I error was estimated as the percentage of inaccurate significance detection for $\beta_2$. The author used the glm function in the HGLM package in R for the original binomial beta h-likelihood method. Using the glm function, an estimation of parameters $\beta$ and $t$-statistics with p-values is obtained. By simulation, an average of 1,000 estimates was calculated for $\beta_1, \beta_2$, power of the hypothesis test for $\beta_1$. Type I error of the hypothesis test for $\beta_2$ and standard error for $\beta_1$. The adjusted h-likelihood is to obtain a different beta distribution scale for a random variable to account for over-dispersion. For the adjusted h-likelihood' binomial-beta HGLM,' the investigator writes the h-likelihood function after changing and uses the maxLik function in the maxLik package for maximum likelihood in the R program. Henningsen and Toomet (2011) explained the maxLik function. By using a loop inside the function to account for the distinct scale that the researcher is adjusting, and by using maxLik to estimate $\beta$.

Table 1 summarizes all results obtained from the simulation for the comparison between the binomial beta estimation method and the adjusted binomial beta estimation method based on point estimation of $\beta_1$; $\beta_2$, Type I error, and Standard Error. The values of the statistical power that we got were very close and equal to one, and are then released from comparing BB and ABB h-likelihood method table.

In the Figs. 1-4, the values of comparing BB and ABB h-likelihood method are explored. Fig. 1 and Fig. 2 show the Type I Error by using Binomial- Beta and adjusted Binomial- Beta h-likelihood. While Fig. 3 and Fig. 4 explain the standard Error for $\beta_1$ by using Binomial- Beta and adjusted Binomial- Beta h-likelihood.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Sample size</th>
<th>Binomial-Beta</th>
<th>Adjusted Binomial-Beta</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type I Error</td>
<td>Standard Error</td>
</tr>
<tr>
<td>K=20</td>
<td>10</td>
<td>0.211</td>
<td>-0.09</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>0.202</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.201</td>
<td>0.003</td>
</tr>
<tr>
<td>K=50</td>
<td>10</td>
<td>0.208</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>0.203</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.198</td>
<td>0.002</td>
</tr>
</tbody>
</table>
4. Results and Discussion

Table 1 shows that the statistical power is approximately one, since unbalanced data is considered an unequal number of data units over the K clusters. Generate distributions of 10, 25, and 100 from Poisson at random. This implies that the amount of observation ‘sample sizes’ for each mix is big, roughly 200 sample sizes for each combination. The size of the sample may have an impact on the power of each technique due to a higher sample size. In all simulations carried out as mentioned above, Estimation of the parameters points, Type I error, and SE of the simulations shown in the past Comparing BB and ABB h-likelihood method and graphs. The statistical power was calculated from the simulation outcomes when the rejected hypothesis $H_0: \beta_1 = 0$ was correct. Calculate by simulation how many times the test is significant for 1000 times. Power is the percentage of the amount rejected correctly calculated. The greater the power, the better the method, is difficult to decide because the power is 1, and high for two techniques because the sample size is big for each mix. It is sensible elevated power for large sample size, there is no difference between two techniques in power, and two techniques operate well by power for large sample size. Methods operate well according to the power of the large sample size. Type I error rates were calculated as the percentage of p values less than 0.05 in the null hypothesis $H_0: \beta_1 = 0$ of no treatment impact when wrongly rejected. The smallest value of Type I error is better in statistics. As indicted in Fig. 1 and Fig. 2 the adjusted binomial beta has the lowest type I error value in all cases, which means the best technique if we decide on the type I error. The SE was calculated as an average of 1000 SEs of the $\beta_1$ estimates. The standard error, which explained in Fig. 3 and Fig. 4 for treatment, is to demonstrate whether or not the efficacy is improving. It is a simulation calculator. The smaller SE represents a smaller variability, or greater accuracy, of the estimation of the parameter (Heo and Leon, 2005).

5. Conclusion

From the above, two techniques are unbiased for the parameters and work better for large clusters. Good to understand that the adjusted binomial beta has produced good outcomes for binary outcomes. The adjusted binomial beta provides a better assessment than the binomial beta technique with information that has over-dispersion.

References