Bias Reduction in Estimating Variance Components of Phytoplankton Existence at Na Thap River Based on Logistics Linear Mixed Models

R Arisanti\(^1\), K A Notodiputro\(^2\), K Sadik\(^2\) and A Lim\(^3\)

\(^1\) Department of Statistics, Padjadjaran University, Indonesia
\(^2\) Department of Statistics, Bogor Agricultural University, Indonesia
\(^3\) Department of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Thailand

Email: almiraza@gmail.com, khairilnotodiputro@gmail.com, kusmansadik@gmail.com, api_45@hotmail.com

Abstract. There are two approaches in estimating variance components, i.e. linearity and integral approaches. However the estimates of variance components produced by both methods are known to be biased. Firth (1993) has introduced parameter estimation for correcting the bias of the maximum likelihood estimates. This method is within the class of linear models, especially the Restricted Maximum Likelihood (REML) method, and the resulting estimator is known as the Firth estimator. In this paper we discuss the bias correction method applied to a logistic linear mixed model in analyzing the existence of Synedra phytoplankton along Na Thap river in Thailand. The Firth adjusted Maximum Likelihood Estimation (MLE) is similar to REML but it shows the characteristic of generalized linear mixed model. We evaluated the Firth adjustment method by means of simulations and the result showed that the unadjusted MLE produced 95% confidence intervals which were narrower when compare to the Firth method. However, the probability coverage of the interval for unadjusted MLE was lower than 95%, whereas for the Firth method the probability coverage is approximately 95%. These results were also consistent with the variance estimation of the Synedra phytoplankton existence. It was shown that the variance estimates of Firth adjusted MLE was lower than the unadjusted MLE.

1. Introduction

Breslow and Clayton [1] have introduced a combination of Linear Mixed Models (LMM) and Generalized Linear Models (GLM) in the form of Generalized Linear Mixed Models (GLMM). If \(X\) and \(Z\) are design matrices, \(b\) and \(\beta\) are the corresponding coefficients then the GLMM models can be described by \(\eta = X\beta + Zb\) with \(b \sim N(0, G), \eta = g(\mu)\) and \(y|b \sim G(\mu, R)\). \(G(\mu, R)\) is a function that follows a specific distribution (not necessarily normal) with mean \(\mu\) and the covariance matrix \(R\). This model has three main components, namely fixed effects (\(\beta\)), random effects (\(b\)), and components of variance (\(\sigma\)). A model of GLM with binary responses can be approximated by logistic linear models [6]. Logistic regression models that involve random explanatory variables is are called Logistic Linear Mixed Models.
In mixed models, we are interested in estimation of variance components [5]. To estimate the variance components, it is common to employ linearity and integral approaches. The linearity method can result in REML (Restricted Maximum Likelihood Estimation) methods by substituting $y$ in the linear mixed model with pseudo variables $y^*$. The integral approach is more widely used since it uses the actual likelihood, hence the goodness of fit can be calculated. However, both methods have disadvantages in which the estimates of variance components are biased. The impact of this bias can be seen on the hypothesis testing and confidence interval produced by the model. If the variance estimation is too small, then the estimates will be inflated and tend to reject the null hypothesis. Moreover, the confidence intervals are also too narrow and the p-values are lower than the actual stated values.

Reference [4] introduced an estimation of parameters to adjust for the bias using maximum likelihood method. In this paper, we discussed estimation of variance components for logistic linear mixed models using the Firth correction methods. We then applied the models to the phytoplankton data, a data collected from Na Thap river in South Thailand.

2. Logistic Linear Mixed Model (LLMM)
In this study, we employed a logistic linear mixed model, especially a logit model with random intercepts to estimate the phytoplankton proportion parameter $p$. Then we choose $r$ sites which consist of $n$ samples for each site. The random intercept model is

$$\text{logit}(p_i) = \eta + b_i$$

where $\text{logit}(p_i) = \ln \left( \frac{p_i}{1-p_i} \right)$ is a canonical link, $\eta$ is an intercept, and $b_i$ is the effect of the $i^{th}$ subject and it is assumed distributed as $N(0, \sigma^2)$. Alternatively, we can use the following notation:

$$\text{logit}(p_i) = \eta + \sigma z_i$$

where $z_i$ is the standardized random effects. Gauss-Hermite quadrature of estimation can be used when the $r$ sites contain relatively small $n$ [3].

Gauss-Hermite Quadrature is basically a numerical integral approximation based on normal distributions [7]. The conditional likelihood and log-likelihood equation for one subject can be written:

$$L(\theta|y, z) = \binom{n}{y} \left( \frac{1}{1 + e^{-(\eta + \sigma z)}} \right)^y \left( 1 - \frac{1}{1 + e^{-(\eta + \sigma z)}} \right)^{n-y} \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{z^2}{2}\right)}$$

$$l(\theta|y, z) = \ln \binom{n}{y} + y(\eta + \sigma z) - n(\eta + \sigma z) - n \ln(1 + e^{-(\eta + \sigma z)}) - \ln \sqrt{2\pi} - \frac{z^2}{2}$$

where $y$ as response vector and $z$ as random effect vector. Defined $h(z, \theta)$ as the negative log-likelihood. To solve the fixed effect, we intergrate the random effects using Gauss-Hermite quadrature:

$$\int e^{-h(z, \theta)} dz$$

The second order of Taylor expansion $h(z, \theta)$ is

$$q(z, \theta) = h(\tilde{z}(\theta), \theta) + \frac{1}{2} h''(\tilde{z}(\theta), \theta)(z - \tilde{z}(\theta))^2$$

$\tilde{z}(\theta)$ maximized the conditional likelihood $\theta$, dan $h''(\tilde{z}(\theta), \theta)$ is the second derivative $h(z, \theta)$ for the random effect which evaluated at $\tilde{z}(\theta) = \partial \tilde{z},$
\[ h''(z, \theta) = \frac{1}{\sigma^2} + \frac{ne^{\eta + \sigma z}}{(1 + e^{\eta + \sigma z})^2} \]

Then,
\[
\int e^{-h(z, \theta)} dz = \int e^{-(h(z, \theta) - q(z, \theta))} e^{-q(z, \theta)} dz
\]
\[
= \frac{\sqrt{2\pi} e^{-h(\hat{z}(\theta), \theta)}}{\sqrt{h''(\hat{z}(\theta), \theta)}} \int e^{-(h(z, \theta) - q(z, \theta))} \frac{h''(\hat{z}(\theta), \theta)}{z} (z - \hat{z}(\theta))^2 dz
\]
\[
= \frac{\sqrt{2\pi} e^{-h(\hat{z}(\theta), \theta)}}{\sqrt{h''(\hat{z}(\theta), \theta)}} \int e^{-\left(\frac{z}{\sqrt{h''(\hat{z}(\theta), \theta)}} + \hat{z}(\theta, \theta) - q\left(\frac{z}{\sqrt{h''(\hat{z}(\theta), \theta)}} + \hat{z}(\theta, \theta)\right)\right)} e^{\frac{z^2}{2}} \frac{1}{\sqrt{2\pi}} dz
\]

Assume that \((x_i, w_i)\) is the abscissa and Gauss-Hermite quadrature weight. Therefore,
\[
\int e^{-h(z, \theta)} dz \approx \frac{\sqrt{2\pi} e^{-h(\hat{z}(\theta), \theta)}}{\sqrt{h''(\hat{z}(\theta), \theta)}} \sum_i e^{-\left(\frac{z}{\sqrt{h''(\hat{z}(\theta), \theta)}} + \hat{z}(\theta, \theta) - q\left(\frac{z}{\sqrt{h''(\hat{z}(\theta), \theta)}} + \hat{z}(\theta, \theta)\right)\right)} \frac{x_i^2}{w_i}
\]

In the initialization step, \(\hat{\eta} = \text{logit} (\hat{p})\) and \(\hat{p} = \sum \frac{y_i}{n_i}\) can be used, and then followed by the gradient and Hessian by Newton-Raphson and Broyden method.

3. Firth Bias Reduction

Reference [4] proposed a technique to obtain estimates having less bias when compared to MLE. This technique combined the advantage of REML and inferences based on the true likelihood. Bias correction is resulted from the score equation, \(S(\theta) = \frac{\delta L(\theta | y)}{\delta \theta}\). For parameter equals to \(\theta_0\), then the expected value of score is zero. Therefore, MLE is the inverse score evaluated at zero:
\[
E_{\theta_0}[S(\theta_0) | y] = 0
\]
\[
\hat{\theta}_{MLE} = S^{-1}(0 | y)
\]

Firth proposed two adjustment. First, the adjustment uses the expected of Hessian, known as Fisher information matrix. The adjustment related to the \(j^{th}\) parameter is
\[
A_{\theta j} = -\frac{1}{2} \text{tr} \left( F^{-1} E \left[ S_{\theta j} (H - SS^T) \right] \right)
\]

where \(F^{-1}\) is the inversed of Fisher information matrix, \(H\) is Hessian matrix, \(S\) is score vector and \(S_{\theta j}\) is the score vector element of parameter \(\theta_j\). The second adjustment is done by modifying the expected adjustment:
\[
A_{\theta j} = -\frac{1}{2} \text{tr} \left( H^{-1} E \left[ S_{\theta j} (H - SS^T) \right] \right)
\]
The double observed adjustment to reduce the expected adjustment is given by

\[ A_{\theta j} = -\frac{1}{2} \text{tr}(H^{-1}E[S_{\theta j}(H - SS^T)]) \]

Let \( S = \sum_{i=1}^{r} S(y_i) \) and \( H = \sum_{i=1}^{r} H(y_i) \). The expected value of observed Firth adjustment is

\[ E_y \left[ \sum_{i=1}^{r} S_{\theta j}(y_i) \left( \sum_{k=1}^{r} H(y_k) - \sum_{k=1}^{r} S(y_k) \sum_{l=1}^{r} S^T(y_l) \right) \right] \]

Because \( y_i \) are independent and because \( E_y[S_{\theta j}(y)] = 0 \), then the expected value is

\[ E_y \left[ \sum_{i=1}^{r} S_{\theta j}(y_i) \left( H(y_i) - S(y_i)S^T(y_i) \right) \right] \]

The double observed adjustment uses the observed information matrix. Then, we use

\[ A_{\theta j} = -\frac{1}{2} \text{tr} \left( H^{-1} \sum_{i=1}^{r} S_{\theta j}(y_i) \left( H(y_i) - S(y_i)S^T(y_i) \right) \right) \]

To obtain the estimates iteratively, the Broyden method can be used.

4. Result and Discussion

In this study, we used survey data which was carried out by the Electricity Generation Authority of Thailand (EGAT) taken from June, 2004 until April, 2014. The data was collected every two months along the Na Thap river, Thailand. In this survey, EGAT divided the Na Thap river into ten sites from three zones as illustrated in Figure 1.

![Figure 1. Site at Na Thap river.](image-url)
Six divisions of phytoplankton have been found in Na Thap river. The number of genus was 53 genus [8]. The divisions are (1) 4 genus of Cyanophyta, (2) 14 genus of Chlorophyta, (3) 21 genus of Bacillariophyta, (4) 4 genus of Pyrrophyta, (5) 2 genus of Euglenophyta and (6) 1 genus of Cryptophyta. There are three dominant divisions, i.e Bacillariophyta, Chlorophyta and Pyrrophyta [2]. We used the Synedra phytoplankton data to evaluate the Firth method.

4.1. The Simulation
A simulation has been carried out to show the performance of Firth method in reducing the variance components. Table 1 showed the simulation results.

|                          | Firth adjusted MLE | Unadjusted MLE |
|--------------------------|--------------------|----------------|
| Standard deviation of sampling | 0.4369             | 0.4352         |
| Standard error mean      | 0.4361             | 0.3960         |
| Estimated coverage probability | 94.73%             | 91.52%         |

The table showed that the standard deviation of the unadjusted MLE (0.4352) was slightly lower than the standard deviation of the Firth adjusted MLE (0.4369). The table has provided evidence that there was bias in variance estimation in the logistic models. Moreover, the standard error of unadjusted MLE for the treatment effects (0.3960) was lower than the standard deviation of the sampling distribution (0.4352). This result produced narrow confidence intervals and low coverage probabilities (91.52%) for the unadjusted MLE. On the other hand, the standard error of the Firth adjusted MLE was 0.4361. This was much closer to the standard deviation of the sampling distribution (0.4369). The Standard error of Firth adjusted resulted in wider confidence intervals than those for the unadjusted MLE, but the coverage probabilities are close to the 95%.

4.2. The Existence of Synedra Phitoplankton at Na Thap River
Now, we will show the application of the Firth adjustment in estimating variance components of Synedra phytoplankton existence in Na Thap river. The application of this model to the phytoplankton data produced estimates of fixed effects and random effects for Synedra existence as shown in Table 2.

|                | Unadjusted MLE | Firth Adjusted |
|----------------|----------------|----------------|
|                | Estimates   | Std. Error | t-stat | Pr(|z|)  | Estimates   | Std. error | t_stat  | Pr(|z|)  |
| Intercept     | -0.5922  | 1.7165    | -0.345  | 0.0071 | -0.6860   | 1.4218     | -0.4825 | 0.0063  |
| DO            | 0.4243   | 0.3112    | 1.363   | 0.0173 | 0.4135    | 0.2330     | 1.7752  | 0.0007  |
| BOD           | -0.3716  | 0.2674    | -1.389  | 0.0165 | -0.3411   | 0.2316     | -1.4728 | 0.0014  |
| Salinity      | 0.0002   | 0.0007    | 0.304   | 0.0076 | 0.0002    | 0.2316     | 0.2816  | 0.0077  |
| \( \sigma^2 \) residual | 2.175    | 1.363     |        |        |           |            |         |        |
The table showed that the variance of residuals using the Firth method was 1.363. This variance was smaller than residual variance of the unadjusted MLE (2.175). It means that the Firth method has reduced the bias of variance components of Synedra existence in the Na Thap river. The table also showed that three predictor variables have affected significantly to the Synedra existence either using unadjusted MLE and Firth adjusted method. The p-values of DO, BOD, and Salinity variables were less than 0.05.

5. Conclusion
We have showed that the mean estimates of the Firth adjusted MLE were closer to the true value than the unadjusted MLE. The Firth adjusted MLE was found to produced 95% probability coverage for the confidence interval whereas the unadjusted MLE produced less than 95% probability coverage. In general we showed that the Firth adjustment is effective to obtain lower variance component, and hence it can be used to correct the underestimated variance components.

Acknowledgements
We thank Dr. Sarawuth Chesoh at Prince of Songkla University in Pattani for giving us permission to use the Na Thap river data as part of the international research collaboration with Bogor Agricultural University.

References
[1] Breslow NE and Clayton DG 1993 *Journal of the American Statistical Association* **88**(421) 925
[2] Chesoh S and A Lim 2015 Monitoring of Thermal Power Plant Operation on Aquatic Ecosystem in Tropical Estuarine River of Thailand, 2008-2013. IAHR World Congress. Netherlands
[3] Classen EA 2014 A Reduced Bias Method of Estimating Variance Component in Generalized Linear Mixed Model Lincoln: University of Nebraska
[4] Firth D 1993 *Biometrika* **80**(1) 27-38
[5] Henderson CR1953 *Biometrics* **9**(2) 226_252
[6] McCullagh P, Nelder JA 1989 Generalized Linear Models, 2nd. New York: Chapman and Hall
[7] Pinheiro JC, Chao EC 2006 *Journal of Computational and Graphical Statistics* **15**(1) 5881
[8] Saheem, N. 2015. Statistical Modelling of Aquatic Animal Abundance in the Na Thap River. A Thesis submitted in fulfillment of the requirements for the Degree of Doctor of Philosophy in Research Methodology. Prince of Songkla University.