IJournals: International Journal of Social Relevance & Concern
ISSN-2347-9698
Volume 4 Issue 9 September 2016

Expectation and Maximization Algorithm for Estimation in Random and Fixed Effect Model in Mixed Model

Adji Achmad Rinaldo Fernandes

Affiliation: Statistics, Faculty of Mathematics and Natural Sciences, University of Brawijaya, Malang, Indonesia

E-mail: fernandes@staff.ub.ac.id

ABSTRACT

Several response variables were categorized only two groups or classes in every subject classified observations on being successful or unsuccessful. Such conditions tend to the binomial distribution. Binomial distribution is often found in the response variables are correlated to longitudinal data. Longitudinal data with binomial distributed response variables can be modeled in Mixed Model. Mixed Model used to model the longitudinal data on clinical research and epidemiological studies such as cancer and other diseases. The purpose of this study was to test whether the algorithm Expectation Maximization (EM) to estimate model parameters Mixed Model better than commonly used algorithms that Newton Rhapson (NR) algorithm. This study uses four simulation of data in health research. Based on the research data, it can be concluded that the EM algorithm to estimate the parameters better than the models Mixed Model NR algorithm.

Keywords: *Mixed Model, EM algorithm, and the algorithm NR*

1. INTRODUCTION

Frequently encountered in the real world that has two possible events, eg healthy or sick, rain or not, and so forth, where the data type is called binary data. In general, binary data is assumed to spread binomial, which is denoted in the form of success (denote by 1), or fail (denote by 0). Longitudinal data is data obtained from repeated measurements (repeated measures) in some individuals (cross-sectional units) within a row (unit of time). Verbekke and Molenberghs [1] introduced the method of analysis of Mixed Model were used in longitudinal data with binomial or binary response.

In the Mixed Model, there are two effects allegedly estimated, the first, the fixed effect which is the effect of the treatment (treatment) and the effect of the concomittant variables, and secondly the random effect which is the effect of the differences between individuals (subject specific). Both methods are expected to approach simultaneous Maximum Likelihood (ML) to estimate the fixed effects, and the Restricted Maximum Likelihood (REML) to estimate the random effects, with the help of Newton Rhapson (NR) iteration approach. Lately, there are some disadvantages that occur in the NR algorithm as the results that appear in the algorithm are the negative range which give rise to inadmissible solution or a solution that is not acceptable.

Meng and Dyk [2] has developed an Expectation Maximization (EM) algorithm that uses a two-step (estimation and maximization), the mixed effect models to the case of quantitative response (interval and ratio scale data). Therefore, in this study it raised the development of EM algorithm in the fixed effect estimation method (fixed effect) by using the ML, and the method of estimation of random effects on the model using REML Mixed Model, which is the sustainability of the research conducted by Meng and Dyk [2], where the difference is, in this study involves binary response variable. In this study wanted at once to compare whether the EM algorithm is better than the NR algorithm, by looking at the value of goodness of fit is the Akaike Information Criterion (AIC) of the two algorithms.

The purpose of the research to be obtained is to estimate the model parameters in Mixed Model for binary response variable using the EM algorithm, as well as, testing the goodness of the EM algorithm with NR, using the smallest AIC value. While the benefits of the research is as an alternative to the settlement issue on longitudinal data analysis with binary response, and the development of the EM algorithm on longitudinal models are expected to be used as the best alternative for the estimation of the model parameters, so it will not happen again in admissible solution.

2. MODEL PROPERTIES

2.1. Mixed Model

Mixed Model is the development of the Generalized Linear Model (GLMS). GLMS models for binary response known as logistic regression. Agresti [3] states, if there is a response variable Yi chances of success for the variables, then:

$$Y_i = \begin{cases} 1, P(Y_i = 1) = \mu_i \\ 0, P(Y_i = 0) = 1 - \mu_i \end{cases}$$

If the number of trials which is denoted by n by 1 then Yi follow the Bernoulli distribution, and if $n \geq 2$ then Yi follow the Binomial Distribution (n_i). So that the logistic model is obtained as follows (Fahrmeir, and Gerhard, [4]):

$$\ln\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$$

$$\mu_i = \frac{\exp(\beta_o + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})}{1 + \exp(\beta_o + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})}$$

$$g(\mu_i) = \ln\left(\frac{\mu_i}{1-\mu_i}\right)$$

$g(\mu_i)$ is the logit link function of the Binomial Distribution

Mixed Model is an extension of Generalized Linear Models (GLMS) for data correlated as in longitudinal data with random effects added to the equations. Response variable in Mixed Model assumed to be independent with the addition of random effects for each subject. In general, the model in Mixed Model is:

$$g(\mu_i) = \mathbf{X}^T \boldsymbol{\beta} + \mathbf{Z}^T b_i + e_i$$

$$b_i \sim N(0,\mathbf{D}), \text{ dan } e_i \sim N(0,\mathbf{R}_i)$$

where :

S_i : i^{th}subject, i=1,2,3,…,N

W_j : j^{th} observation, j=1,2,3,…,n

$g(\mu_i)$: link function(logit)

μ_{ij} : mean of the i^{th}subjectandj^{th}observation

X : covariate vector offixedeffects forithsubjectandjthobservation

β : fixed effectsestimator

Z : covariate matrix of random effect for ithsubjectandjthobservation

b_i : random effect estimator ithsubject

e_i : errormodelforithsubject

D : Variance ofrandomeffects

R_i : Variance of Error Model forithsubject

2.2. Comparative Indicators of NR and EM Algorithm

Selection of the best algorithms by using AIC (Akaike Information Criteria), which is defined in the following equation:

AIC =-2p + 2loglikelihood

p is the number of parameters to be estimated. According to Agresti [3], the best algorithm is the algorithm that produces the smallest AIC value.

3. MODEL IMPLEMENTATION

The data obtained are two primary data from patients with dengue fever, and patients with decubitus wound, and two secondary data of patients Age Related Macular Degeneration (ARMD), and experiments on pregnant mice.

3.1. Newton Rhapson (NR) Algorithm

According to Saavedra [5], the assumption in Mixed Model is:

1. Expected value of the response variable and the covariates associated with the random effects such as:

   $$g(\mu_i) = g[E(Y_{ij}|b_i)] = \mathbf{X}^T \boldsymbol{\beta} + \mathbf{Z}^T b_i$$

2. Assume to each of the subjects that are assumed to be independent, and follow GLMS where y has the density function of the exponential family f (y | β, D, σ2) where D is the covariance matrix.
3. With independent random effects follow a normal distribution, Mixed Model can be used to analyze discrete longitudinal data, including binary response (Hardin and Hilbe, [6]). The relationship between the response variable and the fixed effects parameters as random effects in the following equation:

   $$Y_{ij}|b_i \sim \text{Bernoulli}(\mu_{ij})$$

   $$\ln\left(\frac{\mu_{ij}}{1-\mu_{ii}}\right)_i = \mathbf{X}^T \boldsymbol{\beta} + \mathbf{Z}^T b_i$$

The difference between the above equation with the previous equation is the inclusion of random effects bi in the model, the function of the random effects in the above equation is to address the correlation between each of the observations that may arise in longitudinal data. Random effect is a component of variation that is not explained in the predictor variable. While the fixed effects are caused by the influence of the predictor variables.

According Molenberghs and Verbekke [1] random effects can be predicted by the method of Maximum Likelihood (ML), which is obtained by random effects integrate. According to Jiang [7], the equation for the likelihood of each subject is:

$$L(\beta, D) = \prod_{i=1}^{N} f(y_i|\beta, D) = \prod_{i=1}^{N} \int \prod_{j=1}^{n} f_{ij}(y_{ij}|b_i, \beta) f(b_i|D) db_i$$

To complete the equations in the model is not easy because βj Mixed Model which is nonlinear in parameter, it is necessary for iterative methods. According to Khuri [8], iteration is necessary if optimum value can not be obtained directly. Iterative method used is the NR algorithm. NR approach is generally defined as follows: A point x in a function f (x) are nonlinear approximated using NR method is (Verbekke and Molenberghs, [1]):

$$x^{(t+1)} = x^{(t)} + \left[f''(x^{(t)})\right]^{-1} f'(x^{(t)})$$

where :

$x^{(t+1)}$: resultspointiteration$t$+1

$x^{(t)}$: starting pointora point ofiteration$t$-th

$f'(x)$: the firstderivativeoffunction$f(x)$

$f''(x)$: the secondderivativeoffunction$f(x)$

Analogous to theabove equation, the equation approachesforparameterβis :

$$\beta^{(t+1)} = \beta^{(t)} - (H^{(t)})^{-1} g^{(t)}$$

$$H^{(t)} = \frac{\partial^2 \ell(\beta_j)}{\partial \beta_j \beta_k} \; \frac{-\partial^2 \ell(\beta_j)}{\partial \beta_j \beta_k} = \sum_i \frac{x_{ij} x_{ik} N_i \exp(\sum_i \beta_j x_{ij})}{\left[1+\exp(\sum_i \beta_j x_{ij})\right]^2}$$

$$= \sum_i x_{ij} x_{ik} N_i \pi_i (1-\pi_i) = \mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}$$

$\mathbf{V}^{-1}$ is diagonalmatrixwithelement $N_i \pi_i (1-\pi_i)$

$\mathbf{H}^{-1}$ is covariance matrixwhich isthe inverseof $\mathbf{H}$

$$g^{(t)} = \frac{\partial \ell(\beta_j)}{\partial \beta_j} = \sum_i \sum_j x_i y_i - \sum_i N x_{ij} \pi(x_i)$$

$$= \sum_i x_{ij} (y_i - N_i \pi_i)$$

$$= \mathbf{X}^T (\mathbf{Y}_i - \mu_i)$$

so thefixed effectsestimator

$$\boldsymbol{\beta}_{(r+1)} = \boldsymbol{\beta}_{(r)} + (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} (\mathbf{X}^T (\mathbf{Y}_i - \mu_i))$$

in the same way, fixed effectsestimatoris obtained :

$$\mathbf{b}_{(r+1)} = \mathbf{b}_{(r)} + \mathbf{D}\mathbf{Z}^T \mathbf{V}^{-1} (\mathbf{Y}_i - \mu_i)$$

This equation must be solved iteratively (r is the index for each iteration, with r = 0, 1, 2, ...), the process is repeated until the obtained β and b are convergent or. β is considered convergent if the value of δ is less than 10-6. The relationship between the above two equations lead to the possibility of NR algorithm gives results that are not convergent, and even create a negative range because of the two kinds are used for both equations, ie a wide range of errors and random effects models.

3.2. Expectation Maximization (EM) Algorithm

EM algorithm is essentially as it has been studied by Meng [2] is a refinement of the NR algorithm for the longitudinal model approach to continuous response. In the previous equation, Ri is the variance of the error, using the following approach:

$\mathbf{R_i} = \sigma^2 \mathbf{I}_{ni}$, for i = 1, 2, …, n

where bi and ei ¬ under the assumption of normality is bi distribute to N (0, D), and ei distribute to N (0, Ri), then the Maximum Likelihood method becomes:

$$\hat{\sigma}^2 = N^{-1} \sum e_i^T e_i$$

$$\hat{D} = n^{-1} \sum b_i b_i^T$$

The above equation is M-step (Maximization Step) on the EM-Algorithm, because ei and bi are not known. EM algorithm is a step change with:

$$\sigma^2 = E(\hat{\sigma}^2 \mid y, \beta = \hat{\beta})$$

$$\sigma^2 = N^{-1} \sum \left\{ e_i^T e_i + \sigma^2 \left[n_i - \sigma^2 \mathrm{tr}(V_i^{-1}) \right] \right\}$$

$$D = E(\hat{D} \mid y, \beta = \hat{\beta})$$

$$D = n^{-1} \sum \left\{ b_i b_i^T + \left[D - DZ_i^T V_i^{-1} Z_i D \right] \right\}$$

The right side of the above equation, the variety D is not known, so it was replaced with and the above equation as the initial value. By iteration, and the value will be updated using the above equation by using the EM algorithm approach, so that the EM algorithm scheme are as follows:

Set r = 0, 1, 2, ... is a sequence of iterations, and the expected value of the parameter and is the fixed effects and random effects at iteration r. Parameter estimation steps with the EM algorithm is as follows:

Step 0:

$$\text{set } r = 0,\ \hat{\sigma}^2_{(r)} = 1 \text{ dan } \hat{D}_{(r)} = \mathbf{I}_n$$

Step 1:

set r = r + 1, update equation $\hat{\beta}_{(r)}$ and $\hat{b}_{(r)}$ using:

$$\hat{\beta}_{(r)} = \left(\sum X_i^T \hat{V}_{i(r-1)}^{-1} X_i \right)^{-1} \left(\sum X_i^T \hat{V}_{i(r-1)}^{-1} y_i \right)$$

$$\hat{b}_{(r)} = D_{(r-1)} Z_i^T \hat{V}_{i(r-1)}^{-1} (y_i - X\, \hat{\beta}_{(r)})$$

where

$$\hat{V}_{i(r-1)} = Z_i \hat{D}_{(r-1)} Z_i^T + \hat{\sigma}^2_{(r-1)} I_n$$

Step 2:

recalculate value $\hat{\sigma}^2_{(r)}$ dan $\hat{D}_{(r)}$

$$\hat{\sigma}^2_{(r)} = N^{-1} \sum \left\{ e_{i(r-1)}^T e_{ir-1)} + \hat{\sigma}^2_{(r-1)} \left[n_i - \hat{\sigma}^2_{(r-1)} \mathrm{tr}(\tilde{V}_{i(r-1)}^{-1}) \right] \right\}$$

$$\hat{D}_{(r)} = n^{-1} \sum \left\{ b_{i(r)} b_{ir)}^T + \left[\hat{D}_{(r-1)} - \hat{D}_{(r=1)} Z_i^T \hat{V}_{i(r-1)}^{-1} Z_i \hat{D}_{(r=1)} \right] \right\}$$

where

$$e_{i(r)} = y_i - X_i \hat{\beta}_{(r)} - Z_i \hat{b}_{(r)}$$

Step 3:

$$\text{repeat step 1 and 2 up to convergen } \left| \beta_{j(r+1)} - \beta_{j(r)} \right| \le 10^{-6}$$

3.3. Result of Parameter Estimation

From exploration results it shows that the fourth data are worth to do modeling. Results of Generalized Mixed model predictionfor the first data ie data on patients with dengue fever, are presented as Table 1. From the table below we can see that the time variables significant at α of 5% for both algorithms. Value estimate is positive indicating an increase in healing patients with dengue fever all the time 1 to 4 days. As for the concomitant variables of age, there has been a significant and positive impact on both algorithms. That is, the younger the patient, the higher the cure rate. Seen also in the algorithms for concomitant variables gender seen significant effect on the increase in healing patients with dengue fever. Given the gender is a dummy variable (1: male, and 0: female), indicates a woman has a faster cure rate than men. Overall, the two parameter estimators, using NR and EM showed no significant difference.

Table1ParameterEstimationMixed Modelmodelwithfirst data

Parameter	NR			EM		
	Coefficient	SE	P-value	Coefficient	SE	P-value
Intercept	-4.1526	0.8227	0.0001	-4.1145	0.8121	0.0001
Time	1.5233	0.2243	0.0001	1.5233	0.2243	0.0001
Treatment	0.0134	0.0245	0.5835	0.0134	0.0244	0.6865
Gender	-0.0381	0.4127	0.9264	-0.0381	0.4127	0.9269

Table2Parameter Estimation AlgorithmMixed Modelmodelwithfirst data

Parameter	NR			EM		
	Coefficient	SE	P-value	Coefficient	SE	P-value
Intercept	-7.9810	1.3731	0.0001	-22.3161	9.8084	0.0267
Time	0.4663	0.0410	0.0001	1.4297	0.1795	0.0001
Treatment	0.3773	0.2295	0.0985	0.7373	1.7504	0.6752
Age	0.0367	0.0180	0.0415	0.0757	0.1335	0.5730
Gender	1.7048	0.2435	0.0001	5.6219	1.9092	0.0047

Table3ParameterEstimationAlgorithmNRMixed ModelModelwiththird data

Parameter	NR			EM		
	Coefficient	SE	Pvalue	Coefficient	SE	Pvalue
Intercept	-4.9556	1.2192	0.0001	-3.0394	255.6900	0.9909
Time	0.4024	0.0357	0.0001	0.2540	0.0059	0.0001
Age	0.0144	0.0158	0.3890	1.2234	1.2475	0.3278

Table 4 Parameter Estimation Algorithm Mixed Model Model with fourth Data

Parameter	NR			EM		
	Estimator	SE	Pvalue	Estimator	SE	Pvalue
Intersep	-0.8451	1.2888	0.5120	-4.2533	10.8690	0.6999
Time	0.8751	0.1322	0.0001	4.1998	1.5615	0.0145
Treatment	-0.6082	0.1715	0.0004	-2.7751	1.6630	0.1116
weight	-0.0105	0.0092	0.2529	-0.0537	0.0801	0.5112

Results of Generalized Mixed model predictionfor the second data on patients with decubitus wound as Table 2. From Tables 2, it shows that both algorithms estimate parameters showed considerable differences, both seen from the significance value estimate. In the estimation of the parameters with NR algorithm shows the effect of time, age, and gender of the patient to patient response decubitus wound. Whereas the parameter estimation with the EM algorithm shows only the effect of time and gender of the patient that influence the response of patients with decubitus wound.Mixed Model estimation results for the third entry of patients with Age Related Macular Degeneration (ARMD) as Table 3. From the table below, it shows that both methods of parameter estimation in Mixed Model models showed similar results in estimating the value of significance, but showed different results on the value he suspected. Seen the evolution (changes each time) of the response of patients with Age Related Macular Degeneration (ARMD). To estimate the NR algorithm, shows that from week to week, patients with ARMD can be cured by exp (0.4024) or 1.495 times better than the previous week. As for the estimation using the EM algorithm, it is seen that from week to week, patients with ARMD can be cured by exp (0.2540) or 1.289 times better than the previous week.

Mixed Model estimation results for fourth on the provision of data of betel leaf and hydrogen peroxide in pregnant mice as Table 4. In the table above, it shows that both algorithms estimate parameters showed considerable differences, whether it is seen from significance of estimating value. In the estimation of the parameters with NR algorithm, it shows the influence of time and treatment effect in pregnant mice. Whereas the parameter estimation with the EM algorithm, it shows only the effect of time that influence the increase in pregnant mice.

3.5. Comparison of NR and EM Algorithm

Based on the results of the model parameter estimation Mixed Model, whether it is from significance of estimating value, and the magnitude of predictive value tends to make a difference in the two algorithms are NR and EM Algorithm. To test which is the best algorithm, using the criteria of Akaike Information Criterion (AIC), in which the best algorithm is the algorithm that produces the smallest AIC value. The following table summarizes the AIC value and the number of iterations for each algorithm, and the percent effectiveness.

Table 5 Comparison of results of NR and EM Algorithm

No	Number of Iteration		AIC	
	NR	EM	NR	EM
Data 1	23	22	169.1	163.0
Data 2	41	27	410.2	312.8
Data 3	39	11	1255.8	915.4
Data 4	46	21	113.2	81.9

From the table above it can be seen that the lowest AIC value in the fourth research data generated by the EM algorithm. Also evident is the number of iterations required to reach the EM algorithm to estimate the parameters converging point less than the number of iterations with the NR algorithm on the same data. It can be concluded that the estimation of the model parameters Mixed Model with EM algorithm gives better results than the NR Algorithm.

4. CONCLUSION

From the results of research conducted, it can be concluded as follows: (1) EM algorithm can be used to estimate the model parameters in Mixed Model. This was shown by the formation of the four models in the research data using the EM algorithm, (2) EM algorithm is better than NR which is currently often used in model parameters Mixed Model. This is evident from the four research data, the EM algorithm AIC value is smaller than the value of AIC of NR algorithm. From the results of this study, it can be suggested some of the following: (1) EM algorithm can be used as the settlement of the problem in longitudinal data analysis with binary response, and tend to improve model accuracy better than the NR algorithm is now more frequently used. So it can be suggested that EM development is the best alternative models parameters estimating Mixed Model. (2) On further research it is recommended to use the Monte Carlo Markov Chain (MCMC) criterion which helps to define the selection of the best algorithm. It is recommended to study the development of the EM algorithm to the analysis of longitudinal data with ordinal response and response poison.

5. ACKNOWLEDGMENTS

The author would like to thanks to University of Brawijaya for financial support.

6. REFERENCES

[1] Molenbergs.G., dan Verbekke, G., 2005. **Model for Discrete Longitudinal Data**. Springer Series in statistics. New –York:Springer –Verlag.

[2] Meng, X, dan Dyk, D. 2008. Fast EM Type Implementations for Mixed Effect Models. *Journal of the Royal Statistical Society*. Series B. Vol. 60, No. 3. Pp 559-578.

[3] Agresti, A. 2002. **An Introduction to Categorical Data Analysis.** John Wiley & Sons. New York.

[4] Fahrmeir, L. dan T. Gerhard. 1994 . **Multivariate Statistical Modelling Based on Generalized Linier Models**. John Willey dan Sons , New York.

[5] Saavedra, P.A.T. 2006. **Percentile Curves In Binary Longitudinal Data.** http://grad.uprm.edu/tesis/torressaavedra.pdf. Access January, 18th, 2016.

[6] Hardin, J. W. dan J. Hilbe. 2007. **Generalized Linear Models and Extensions**. Stata Press. Texas

[7] Jiang, J. 2007. **Linear and Generalized Linear Mixed Models and Their Application.** Springer Series in Statistics. New York.

[8] Khuri, A. 2003. **Advanced Calculus with Application In Statistics.** John Wiley and Son. New Jersey.