Estimation in Generalized Linear Models with Heterogeneous Random Effects

Woncheol Jang* Johan Lim†

May 19, 2004

Abstract

The penalized quasi-likelihood (PQL) approach is the most common estimation procedure for the generalized linear mixed model (GLMM). However, it has been noticed that the PQL tends to underestimate variance components as well as regression coefficients in the previous literature. In this paper, we numerically show that the biases of the variance components are systematically related to the biases of the regression coefficient estimates, and also show that the biases of the variance components estimates of the PQL increase as random effects become more heterogeneous.

Keywords and Phrases: Generalized linear mixed models; heterogeneity; penalized quasi-likelihood estimator; variance components.

1 Introduction

The generalized linear mixed effects model (GLMM) has been widely used in biometry and medical studies where random effects explain subject specific variations. For example, the heritability or the genetic correlation in the GLMM can be represented as a function of variance components: parameters related to the random effects of the GLMM. Thus, the estimation of

*Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, wjang@stat.cmu.edu

†Department of Statistics, Texas A & M University, College Station, TX 77843-3143, johanlim@stat.tamu.edu

the variance components are of great interest as well as that of regression coefficients.

In most of random effects models including the GLMM, exact likelihood functions involve intractable high–dimensional integrations and are hard to compute. Accordingly, several approximations to the likelihood functions have been proposed in the previous literature. Among many of them, the penalized quasi-likelihood (PQL) by Breslow and Clayton (1993) is the most popular. It approximates the high-dimensional integrations using the well–known Laplace approximation and the approximated likelihood functions have those of Gaussian distributions. Subsequently, it suggests to use the similar iterative numerical procedures introduced in Harville (1977) to maximize the PQL. It should be noted that the Laplace approximation, in fact, is the most simple approximation procedure for multiple integrations and there are various estimators equivalent to the PQL in the literature including Schall (1991), Wolfinger (1993) and McGilchrist (1994). Especially, Wolfinger's algorithm becomes the basic routine for the SAS procedure.

Even though the PQL is widely used in many different applications, it has been noticed that estimating the variance components is quite challenging due to their unobservability. Accordingly, several different types of the likelihood functions of the variance components have been proposed including the restricted maximum likelihood (REML) estimator by Harville (1977) and the maximum adjusted profile h-likelihood (MAPHL) estimator by Lee and Nelder (1996). However, most of such variance components estimators have not got as much attention as regression coefficients estimators have. Furthermore, they have not been understood well when observations are from non-Gaussian distributions.

In this paper, we numerically study the performance of the PQL estimates along with the REML estimates of the variance components which are the most commonly used in practice. In the remainder of the paper, the PQL estimates assumes that the REML estimates are used for the variance components.

This paper is composed of two folds. First, we investigate the performance of the PQL estimates and show that the biases of the variance components estimate results in systematic biases of the the regression coefficient estimates.

A simulation study shows that the PQL underestimates both the regression coefficients and variance components in the GLMM, and also shows that the biases of the regression coefficient are closely related to those of the variance components estimates. Here, the conjectured close relationship between the biases of the estimates of the regression coefficients and those of the variance components is consistent with the previous study stating that the regression coefficient estimate is biased downward to 0 if the random effects are mistakenly ignored (Neuhaus , 1998; Henderson and Oman , 1999).

Second, we investigate the performance of the PQL estimates in the GLMM when the random effects are heterogeneous. Here, heterogeneous random effects imply the distributions of random effects are different from one group of subjects to the other group of subjects. Such heterogeneous random effects have recently reported and studied by Qiou et al (1999). The second simulation study shows that the biases of the variance components estimates by the PQL as well as those of the regression coefficient estimates increase as the random effects become more heterogeneous.

This paper is organized as follows. Section 1 is a brief introduction. In Section 2, we review the GLMM and the PQL. Section 3 implements two simulation studies on the performance of the PQL estimators. Finally, Section 4 provides discussion and summarizing our results.

2 Models and Methods

In this section, we review the GLMM and the PQL by Harville (1977) and Breslow and Clayton (1993).

2.1 Generalized Linear Mixed Models

The generalized linear models with random effects have been studied over two decades and are the most common models for longitudinal data analysis in recent studies.

Suppose that the repeated observations $y_{i,1}, y_{i,2}, \cdots, y_{i,n_i}$ of the ith subject are observed along with the covariates $x_{i,1}, x_{i,2}, \cdots, x_{i,n_i}$ (each $x_{i,j}$ is a $p \times 1$ vector) for fixed effects and $z_{i,1}, z_{i,2}, \cdots, z_{i,n_i}$ (each $z_{i,j}$ is a $q \times 1$ vector) for

random effects. Here, N subjects are assumed.

Let the unobserved random effects be $b = (b_1, \cdots, b_N)$, where each b_i is a $q \times 1$ vector. Then, it is assumed that, given b, y_{ij}s are independent of each other and are from a distribution with mean and variance as:

$$\mathrm{E}(y_{ij}|b) = \mu_{ij}^b \quad \text{and} \quad \mathrm{Var}(y_{ij} \mid b) = \phi a_{ij}^{-1} V(\mu_{ij}^b), \tag{1}$$

where ϕ is a disperson parameter, a_{ij} is a prior weight and $\mathrm{V}(\cdot)$ is a variance function, subjectively specified. Further, the link function in the GLMM is specified as

$$g(\mu_{ij}^b) = x_{ij}^T \alpha + z_{ij}^T b. \tag{2}$$

Equivalently, it can be expressed as using a matrix notation as:

$$g(\mu_i^b) = X_i^T \alpha + Z_i^T b. \tag{3}$$

where $\mu_i^b = (\mu_{i,1}, \ldots, \mu_{i,n_i}^b)^T$ and the design matrix X_i and Z_i have rows $x_{i,j}^T$ and $z_{i,j}^T$. Here, α is a $q \times 1$ vector of the fixed effects and the random effects b follow a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $D = D(\theta)$, and θ is a $c \times 1$ unknown vector of the variance components. In this paper, we assume that the disperson parameter ϕ and the prior weight $a_{i,j}$ are unities and $D = \mathrm{diag}(\theta_s \oplus I_{q_s})$ for $s = 1, \ldots, c$ and $\sum_{s=1}^c q_s = q$. In other words, we assume that the random effects are independent to each other.

2.2 Penalized Quasi-Likelihood Method

Let y_{ij}^* be a linearized form of the link function which is specified as

$$y_{ij}^* = g(y_{ij}) = g(\mu_{ij}^b) + (y_{ij} - \mu_{ij}^b) g'(\mu_{ij}^b).$$

Then, based on the original model, one can describe the distribution of y_{ij}^* as a linear model with structure,

$$Y_i^* = X_i \alpha + Z_i b + \epsilon_i,$$

where

$$Y_i^* = (y_{i,1}^*, \ldots, y_{i,n_i}^*)^T, \quad \epsilon_i \sim N(\mathbf{0}, W_i^{-1}),$$

and W_i is the diagonal matrix of $w_{i,1}, \ldots, w_{i,n_i}$, where

$$w_{i,j} = \{V(\mu_{i,j}^b)(g'(\mu_{i,j}^b))^2\}^{-1}, \quad \text{for } j = 1, \ldots, n_i, \quad i = 1, \ldots, N.$$

It is then natural to extend Harville's approach to this setting. We briefly summarize the procedures as follows:

Step 1 Given θ, α and b, one can estimate the fixed effect α by solving the normal equation

$$\sum_{i=1}^{N} X_i^T V_i^{-1} X_i \alpha = \sum_{i=1}^{N} X_i^T V_i^{-1} Y_i^*,$$

where $V_i = W_i^{-1} + Z_i D Z_i^T$.

Step 2 The random effects b can be estimated as

$$\hat{b} = \sum_{i=1}^{N} D Z_i^T V_i^{-1} (Y_i^* - X_i \widehat{\alpha}).$$

Step 3 Subsequently, the REML estimators for θ are

$$\widehat{\theta_s} = \frac{\sum_{n \in Q_s} \widehat{b}_n^2}{\sum_{n \in Q_s} (1 - t_{nn})}, \quad \text{for } s = 1, \ldots, c,$$

where

$$\begin{aligned} Q_s &= \{n : \sum_{i=1}^{s-1} q_i < n \leq \sum_{i=1}^{s} q_i\}, \quad S = W - WX(X^T W X)^{-1} X^T W \\ X^T &= (X_1,^T, \ldots, X_N^T), \quad Z^T = (Z_1^T, \ldots, Z_N^T), \quad W = \text{diag}(W_1, W_2, \cdots, W_N), \end{aligned}$$

and t_{nn} is the nth diagonal element of $T = (I + Z^T S Z D)^{-1}$.

Step 4 One then updates Y_i^* at the end of each iteration. The PQL estimators are defined to be those upon convergence.

Finally, the covariance matrix of the estimates can be computed at the value $\alpha = \widehat{\alpha}$ and $b = \widehat{b}$ by

$$\text{Cov}(\hat{\alpha}) = \Big\{ \sum_{i=1}^{N} X_i^T V_i^{-1} X_i \Big\}^{-1}, \quad \text{Cov}(\hat{\theta}) = H^{-1}.$$

Here H has components

$$h_{st} = \Big[\frac{1}{2}\sum_{i\in Q_s}\sum_{j\in Q_t}\Big(Z_{(i)}^T P Z_{(j)}\Big)^2\Big],$$

where

$$\begin{aligned} Z_{i,n}^T &= (z_{i,1n},\ldots,z_{in_in}), \quad Z_{(n)}^T = (Z_{1n}^T,\ldots,Z_{Nn}^T), \\ V^{-1} &= \text{diag}(V_1^{-1},\cdots,V_N^{-1}), \quad P = V^{-1} - V^{-1}X\text{Cov}(\hat{\alpha})X^T V^{-1}. \end{aligned}$$

More details on this section can be referred from Harville (1977) and Breslow and Clayton (1993).

3 Simulation Studies

In this section, we implemented two simulation studies to investigate the performance of the PQL estimates in the GLMM. First, we study how the regression coefficient estimates vary according to the magnitude of the biases of the variance components estimates. Second, we evaluate the performance of the estimates by the PQL for heterogeneous random effects at various levels of the heterogeneity.

In both simulations, simple logistic regressions with random intercepts are used. Each simulated data set has 50 subjects with 4 repetitions in each subject. Bernoulli random variables, y_{ij}s, are generated at each subject with conditional mean μ_{ij}^b given by

$$\log\Big(\frac{\mu_{ij}^b}{1-\mu_{ij}^b}\Big) = b_i + \alpha_0 + \alpha_1 x_{ij1} + \alpha_2 x_{ij2},$$

where x_{ij1} and x_{ij2} are independently from $N(0,1)$ for $i = 1,\ldots,50$ and $j = 1,\ldots,4$..

The fixed effects are set to be $\alpha^T = (0.5, 2.0, 0.0)$ and the random effects b are generated from a multivariate normal distribution with mean $\mathbf{0}$ and covariance $D = \text{diag}(\theta_1,\theta_2) \oplus I_{25}$. In other words, the first 25 subjects would have random intercepts with variance θ_1 while those in the second half of the subjects have random intercepts with variance θ_2.

θ	$\alpha_0(=0.5)$	$\alpha_1(=2.0)$	$\alpha_2(=0.0)$
0.1	0.4456 (0.2166)	1.7695 (0.2730)	-0.0038 (0.1766)
0.5	0.4438 (0.2286)	1.7959 (0.2617)	-0.0057 (0.1821)
1.0	0.4602 (0.2238)	1.8772 (0.2728)	-0.0030 (0.1896)
1.5	0.4657 (0.2304)	1.9430 (0.2945)	0.0076 (0.2123)
2.0	0.4959 (0.2397)	1.9916 (0.2837)	0.0001 (0.2031)

Table 1: Average of fixed effects and variance components estimates over 500 hundreds data sets; the numbers in parenthesis are the standard deviation of the estimates.

3.1 Downward Bias of the PQL estimates

To show the effects of the variance components estimates to the regression coefficient estimates, we computed the PQL estimates with fixed values of the variance components. To be specific, five hundreds data sets of 200 observations (50 clusters with 4 repetitions) were generated under the above settings with $\theta_1 = \theta_2 = 1.0$ (true value). The PQL estimates are computed after fixing the variance component as constants 0.1, 0.5, 1.0, 1.5, and 2.0. In the estimation process, we used the following initial values $\alpha_1 = \alpha_2 = \alpha_3 = 0, b_i = 0$ for $i = 1, \ldots, 50$.

Table 1 shows the estimates of the regression coefficient estimates for different values of $\theta(= \theta_1 = \theta_2)$. Two interesting observations come from Table 1. First, it can be found that even we set the variance components as its' true value (= 1.0), the PQL estimates of the regression coefficient are still underestimated, in absolute values. This indicates the downward biases of the regression coefficient estimates are not exclusively from the underestimated variance components estimates. Second, it can be found that, as the fixed values of the variance components decrease, the regression coefficients estimates α_1 also decrease toward 0. As pointed out in Section 1, this observations are compatible to the well known results that the regression coefficients are downward biased to 0 when the random effects are mistakenly disregarded (Neuhaus , 1998; Henderson and Oman , 1999).

Scenario	$\alpha_0(=0.5)$	$\alpha_1(=2.0)$	$\alpha_2(=0.0)$	$\theta_1(=0.5)$	θ_2
1	0.4971	1.9488	-0.0037	0.4585	0.4782(0.5)
2	0.4824	1.8972	-0.0063	0.4257	0.7881(1.0)
3	0.4716	1.8600	-0.0049	0.4051	1.0958(1.5)
4	0.4653	1.8311	-0.0045	0.3879	1.3863(2.0)

Table 2: Average of fixed effects and variance components estimates over 100 data sets

3.2 Bias in Heterogeneous Random Effects

To see the effect of the heterogeneous random effects, one thousand data sets of 200 observations were generated with the same data structure (50 subjects with 4 replications). Simulation study implemented under the following 4 scenarios;

Scenario 1: $\theta = (0.5, 0.5)$;

Scenario 2: $\theta = (0.5, 1.0)$;

Scenario 3: $\theta = (0.5, 1.5)$;

Scenario 4: $\theta = (0.5, 2.0)$.

Table 2 contains the average of the estimates of the fixed effects and variance components. Each row in Table 2 shows the average for the corresponding scenarios. The true parameter values for $\alpha_0, \alpha_1, \alpha_2$ and θ_1 are reported inside the parentheses next to the parameters while the true value of θ_2 for each scenario is reported inside the parentheses after the estimates of θ_2.

Using the estimates of α_1 as a typical example of the estimated fixed effects, one may find that the percentage of the biases of $\hat{\alpha}_1$s is 2.6% in senario 1, 5.1% in senario 2, 7% in scenario 3, and 8.4% in senario 4. In other wrods, the larger θ_2 is , the larger the biases of the fixed effects estimates are. In the variance components estimates, the percentage of biases of $\hat{\theta}_2$s dramatically increases from 0.4% to 30.7% as θ_2 increases. The effect of increasing θ_2 on the direction of the biases of $\hat{\theta}_1$s is almost the same as on the direction of

Scenario		α_0	α_1	α_2	θ_1	θ_2
1	Est.	0.2125	0.2939	0.1933	0.6396	0.6482
	Monte.	0.2215	0.3045	0.1939	0.4930	0.5113
2	Est.	0.2164	0.2886	0.1937	0.6207	0.7509
	Monte.	0.2248	0.2996	0.1897	0.4662	0.6681
3	Est.	0.2202	0.2850	0.1942	0.6089	0.8517
	Monte.	0.2298	0.2963	0.1890	0.4504	0.7945
4	Est.	0.2236	0.2823	0.1950	0.5987	0.9486
	Monte.	0.2330	0.2866	0.1891	0.4319	0.9101

Table 3: Comparisons of Estimated and Monte Carlo simulated Standard Errors

those of $\hat{\theta}_2$s, but the percentage of the biases of $\hat{\theta}_1$s are much smaller than those of $\hat{\theta}_2$s as θ_2 increases.

Table 3 shows the comparisons of the estimated and the Monte Carlo simulated standard errors. The "Est." and "Monte" correspond to the estimated and the Monte Carlo simulated standard errors. In the fixed effects, the Monte Carlo simulated standard errors are slightly higher than the estimated standard errors which are derived from the information matrix. The ratios of the Monte Carlo simulated and the estimated variance estimators are almost 1 for all scenarios. However, for the variance components, the Monte Carlo simulated standard errors are much smaller than the estimated standard errors. One may note that the difference between the Monte Carlo and the estimated standard errors of $\hat{\theta}_2$s decreases as θ_2 increases.

Figure 1 and Figure 2 show the distributions of $\hat{\theta}$s. In Figure 1, two boxplots were plotted for $\hat{\theta}_1$ (bottom) and $\hat{\theta}_2$ (top) for each scenario. The true parameter values are indicated in each plot. Figure 2 is similar to Figure 1 except that the boxplots were replaced with corresponding kernel density estimators.

We also observed the long right tails in all $\hat{\theta}$s which is consistent with Lin (1997). Lin pointed out that $\hat{\theta}$ is not exactly normally distributed, unless that the number of clusters is really large and the θs are bounded away form the boundary, 0.

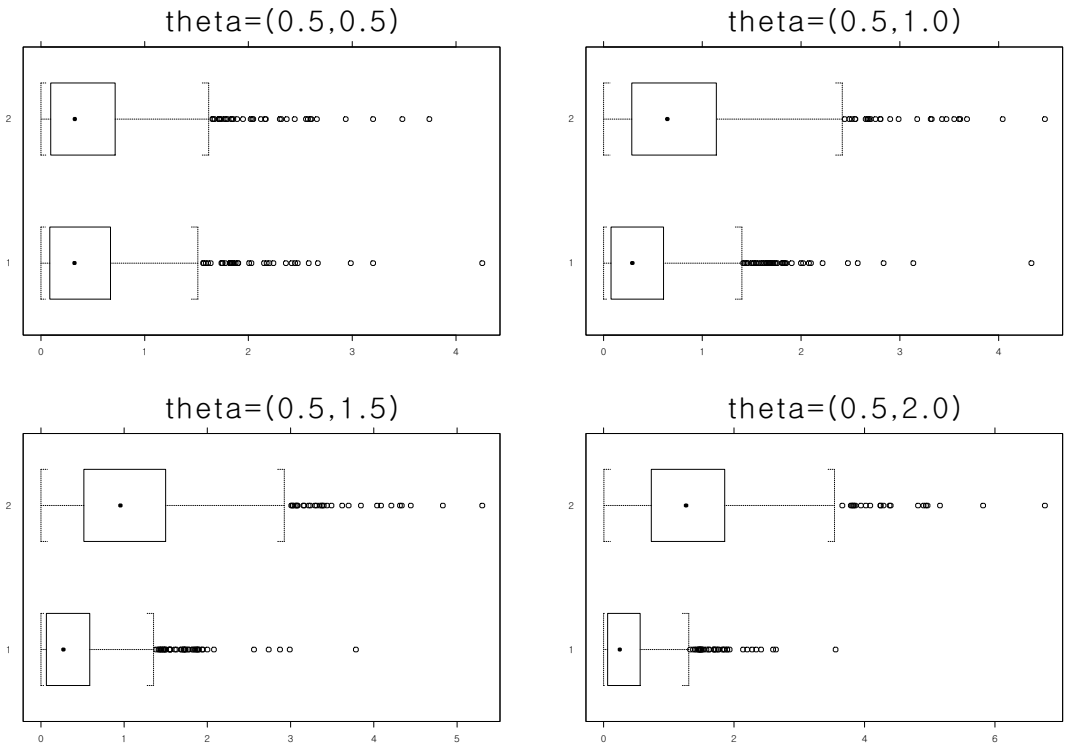

Figure 1: Boxplots $\hat{\theta}$

4 Discussion

In this paper, two simulations studies are implemented to investigate the performance of the PQL estimates. The first simulation study shows that the biases of regression coefficients estimates increase as those of the variance components increase. Also, it reassures that the PQL underestimates both the regression coefficients and the variance components which are briefly pointed out in Breslow and Clayton (1993). Second simulation shows that, using the PQL, the variance components were underestimated while the standard errors of the variance components were overestimated when the random effects are heterogeneous. We also find that regression coefficients are underestimated.

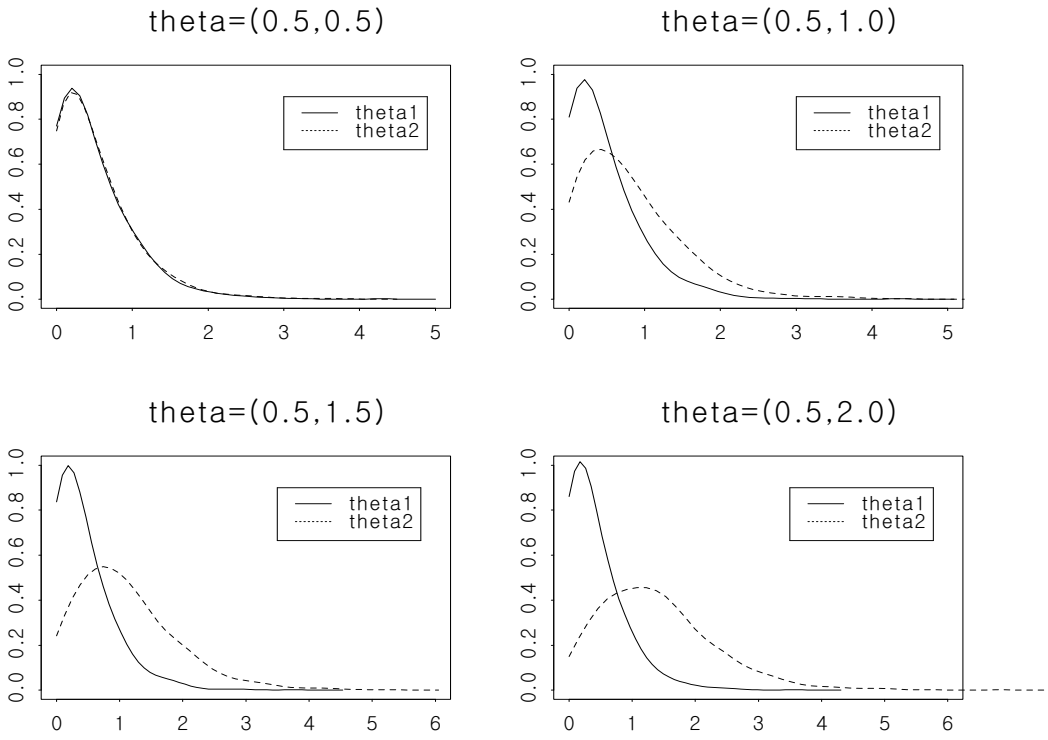

Figure 2: Distributions of $\hat{\theta}$

References

Breslow, N.E. and Clayton, D.G. (1993). Approximate Inference in Generalized Linear Mixed Models. *J. Amer. Statist. Assoc.* **88**, 9-25.

Harville, D.A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *J. Amer. Statist. Assoc.* **72**, 320-340.

Henderson, R. and Oman, P. (1999). Effect of frailty on marginal regression estimates in survival analysis. *J. Roy. Stat. Soc. B* **61**, 367-379.

Lee, Y. and Nelder, J.A. (1996), Hierarchical generalized linear models. *J. Roy. Stat. Soc. B* **58**, 619-6778.

Lin, X. (1997) Variance component testing in generalised linear models with random effects. *Biometrika* **84**, 309-326.

McGilchrist, C.A. (1994). Estimation in Generalized Mixed Models. *J. Roy. Stat. Soc.* **B.** 56:61-69.

Neuhaus, J.M. (1998). Estimation efficiency with omitted covariates in generalized linear models. *J. Amer. Statist. Assoc.* **93**, 1124-1129.

Qiou, Z, Ravishanker, N., and Dey, D. (1999), Multivariate Survival Analysis with Positive Stable Frailties. *Biometrics* **55** 637-644.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika* **78**, 719-727.

Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika* **80**, 791-795.