Communications for Statistical Applications and Methods
2019, Vol. 26, No. 3, 315–323

https://doi.org/10.29220/CSAM.2019.26.3.315
Print ISSN 2287-7843 / Online ISSN 2383-4757

A computational note on maximum likelihood estimation in random effects panel probit model

Seung-Chun Lee[1,a]

[a]Department of Applied Statistics, Hanshin University, Korea

Abstract

Panel data sets have recently been developed in various areas, and many recent studies have analyzed panel, or longitudinal data sets. Often a dichotomous dependent variable occur in survival analysis, biomedical and epidemiological studies that is analyzed by a generalized linear mixed effects model (GLMM). The most common estimation method for the binary panel data may be the maximum likelihood (ML). Many statistical packages provide ML estimates; however, the estimates are computed from numerically approximated likelihood function. For instance, R packages, **pglm** (Croissant, 2017) approximate the likelihood function by the Gauss–Hermite quadratures, while **Rchoice** (Sarrias, *Journal of Statistical Software*, **74**, 1–31, 2016) use a Monte Carlo integration method for the approximation. As a result, it can be observed that different packages give different results because of different numerical computation methods. In this note, we discuss the pros and cons of numerical methods compared with the exact computation method.

Keywords: GLMM, panel regression, Gauss–Hermite quadrature, Monte Carlo integration

1. Introduction

The panel regression model with individual specific effects has the following specification:

$$w_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \epsilon_{it}, \quad i = 1, 2, \ldots, n;\ t = 1, 2, \ldots, T_i, \tag{1.1}$$

where w_{it} is the dependent variable, $\mathbf{x}_{it}$ is the $p\times1$ vector of predictors, $\boldsymbol{\beta}$ is the $p\times1$ vector of regression coefficients, α_i is the time–invariant individual specific effect, and ϵ_{it} is the remaining disturbance term. Here the subscripts i and t represent the individual and the time period, respectively. When α_i is assumed to be constant over time, the model is referred to as the "fixed effects" model, which is known to have incidental parameter problem (Lancaster, 2000), while "random effects" model treats α_i as a random variable.

The probit-normal panel regression model assumes that w is a latent variable and the observed value of dichotomous variable is determined by $y_{it} = 1(w_{it} \geq 0)$, were $1(\cdot)$ is the indicator function, and ϵ_{it}'s are independently and identically distributed standard normal random variables. As McCulloch (1996) stated, a frequentists decision to regard an effect as fixed or random is complicated one, but we will assume that the individual specific effect α_i's are random, and are independent of error term, because the treatment of fixed effects probit-normal panel regression model is the same as the ordinal probit model. Thus, we assume α_i's to be an independently and identically distributed normal random variables with mean 0 and variance σ_α^2. Then, the panel probit regression model is a GLMM.

[1] Department of Applied Statistics, Hanshin University, 137 Hanshindae-gil, Osan, Gyeonggi-Do 18101, Korea.
E-mail: seung@hs.ac.kr

Published 31 May 2019 / journal homepage: http://csam.or.kr

The likelihood function of the GLMM is often quite complex in that it includes multidimensional integrals; in addition, it is also difficult to maximize the likelihood function directly. Many authors proposed to use an integral approximation method for the computation of likelihood function, and then maximize the approximated likelihood function. Various integral approximation methods, such as Gauss–Hermite quadratures, Markov chain Monte Carlo, have been employed for this purpose. These approximation methods enable the ML estimation. Theoretically, these approximations can be improved to any precision; however, such integral approximation methods are inadequate for high dimensional integrals in practice. It may be difficult to get a good approximation when T_i is large. In addition, it is observable that statistical packages using different methods give quite different results. See Zhang *et al.* (2006) for example. Another approach is to linearize the model by the Taylor series, and then maximize the pseudo–likelihood of linearly approximated model. The advantage of this approach is that it is simple so that it can be apply to a model with a large number of random effects or crossed random effects, but it is also known that it gives biased estimates, especially for binary data when the number of observations per subject is small. See Wolfinger and O'Connell (1993) for further details.

Currently, there are many statistical packages to get the ML estimate of a GLMM model. Most of these packages employ the first approach. For instance, **R** packages, **pglm** (Croissant, 2017), **Rchoice** (Sarrias, 2016), **lme4** (Bates *et al.*, 2015) and **nmle** (Pinheiro *et al.*, 2015) and **xtoprobit** of stata approximate the likelihood function either by the Gauss–Hermite quadratures or the Monte Carlo integration method. See Zhang *et al.* (2006) for additional R packages for GLMM. In particular, the first two R packages are specialized to fit the panel probit model. Only a few packages, such as **GLIMMIX** procedure in SAS, adopt the second approach.

This paper demonstrate the effect of approximation on the precision of ML estimate and its standard error. For this purpose, we first calculate the ML estimate base on theoretical log-likelihood function, and then compare this ML estimate with the results of **pglm** and **Rchoice**, which get the ML estimates based on two different approximation methods. Since, **lme4**, **nmle** and other R packages approximate the likelihood function by essentially the same manner of **pglm** or **Rchoice**, we do not examine the other **R** packages. Commercial software were not examined because of license problems; however, **xtoprobit** and **Rchoice** would give similar estimates because they employ the same algorithm.

2. Maximum likelihood method

In this section, we wish to discuss the computation of ML estimate in the random effects panel probit model. Because the theory of ML in the GLMM is quite standard, most of this section may be well known, but we include this section for completeness.

2.1. Estimation

Let $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \sigma_\alpha^2 \mathbf{I})$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ where $\boldsymbol{\alpha}$ and $\boldsymbol{\epsilon}$ are independent random vectors of α_i's and ϵ_{it}'s, respectively, and $N = \sum_{i=1}^{n} T_i$. Then, (1.1) can be written as $\mathbf{w} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is the $N \times p$ matrix of regressors and $\mathbf{Z}$ is the $N \times n$ incident matrix for α_i's. Then, $\mathbf{w}|\mathbf{y}$ is a multivariate truncated normal random vector. Notations related to a multivariate truncated normal distribution are defined as follows. Suppose $\mathbf{z}_{n\times 1} \sim \mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma})$ and $R = \{(z_1, \ldots, z_n) : a_i < z_i < b_i, i = 1, \ldots, n\}$, then the distribution of $\mathbf{z}|\mathbf{z} \in R$ is a multivariate truncated normal on R, and it will be represented by $\mathbf{z}|\mathbf{z} \in R \sim TN_{(\mathbf{a},\mathbf{b})}(\mathbf{m}, \boldsymbol{\Sigma})$, where $\mathbf{a} = \{a_i\}_{i=1}^n$ and $\mathbf{b} = \{b_i\}_{i=1}^n$.

Each element of $\mathbf{y}$ is either 0 or 1. $y_{it} = 0$ indicates that w_{it} is left-truncated, and then define $a_{it}^* =$

$-\infty$ and $b^*_{it} = 0$. Similarly, if $y_{it} = 1$, let $a^*_{it} = 0$ and $b^*_{it} = \infty$. Then, we have $\mathbf{w}|\mathbf{y} \sim TN_{(\mathbf{a}^*,\mathbf{b}^*)}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$, where $\mathbf{V} = \mathbf{Z}\mathbf{Z}'\sigma^2_\alpha + \mathbf{I}$. In particular, if $\mathbf{w}_i$ and $\mathbf{X}_i$, $i = 1, \ldots, n$ represent the partitions of $\mathbf{w}$ and $\mathbf{X}$, respectively, according to each individual, then $\mathbf{w}_i|\mathbf{y}_i \sim TN_{(\mathbf{a}^*_i,\mathbf{b}^*_i)}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i)$, $i = 1, \ldots, n$, where $\mathbf{a}^*_i$ and $\mathbf{b}^*_i$ are partitions of $\mathbf{a}^*$ and $\mathbf{b}^*$, and $\mathbf{V}_i = \mathbf{J}_i\sigma^2_\alpha + \mathbf{I}$. Here, $\mathbf{J}_i$ is a $T_i \times T_i$ matrix of 1's. With these notations, the likelihood function can be written as:

$$L\left(\boldsymbol{\beta}, \sigma^2_\alpha; \mathbf{y}\right) = \int_{\mathcal{R}} \boldsymbol{\phi}(\mathbf{w}; \mathbf{X}\boldsymbol{\beta}, \mathbf{V}) d\mu = \prod_{i=1}^{n} \int_{\mathcal{R}_i} \boldsymbol{\phi}(\mathbf{w}; \mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i) d\mu, \tag{2.1}$$

where $\boldsymbol{\phi}(\cdot\,; \mathbf{m}, \boldsymbol{\Sigma})$ denotes the probability density function of a multivariate normal random vector with mean $\mathbf{m}$ and variance-covariance $\boldsymbol{\Sigma}$, $\mathcal{R} = \{\mathbf{w} : \mathbf{y}(\mathbf{w}) = \mathbf{y}\} = \{\mathbf{w} : \mathbf{a}^* < \mathbf{w} < \mathbf{b}^*\}$ is the set of latent variables given the observed data, and μ is the Lebesgue measure. $\mathcal{R}_i$'s are also defined similarly.

To maximize the likelihood, it needs to evaluate multidimensional integrals. **pglm** (Croissant, 2017) approximates the likelihood by the Gauss-Hermite quadrature method. An alternative approach is to use a simulated likelihood function, an approximation of likelihood by Monte Carlo integration. There were a number issues relating the sampling methods of Monte Carlo integration, see McFadden and Ruud (1994) for example, but **Rchoice** use a small number, say 40 by default, of Halton draws (Halton, 1964) to compute the simulated likelihood function. The approximated likelihood function is then maximized by BFGS (Byrd *et al.*, 1995) gradient methods to get the ML estimate. However, these numerical methods may not be suitable for multidimensional integrals. Since the dimensionality is increasing with T_i, it would be desirable to know the limitation of these algorithms.

The theoretical ML estimate is the solution of likelihood equations obtained by differentiating the log of (2.1) with respect to $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2_\alpha)$, which are

$$S_1(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \log \int_{\mathcal{R}} f(\mathbf{w}) d\mu = \frac{\int_{\mathcal{R}} \mathbf{X}'\mathbf{V}^{-1}(\mathbf{w} - \mathbf{X}\boldsymbol{\beta}) f(\mathbf{w}) d\mu}{\int_{\mathcal{R}} f(\mathbf{w}) d\mu} = \mathbf{X}'\mathbf{V}^{-1}(\mathrm{E}(\mathbf{w}|\mathbf{y}) - \mathbf{X}\boldsymbol{\beta}), \tag{2.2}$$

$$\begin{aligned} S_2(\boldsymbol{\theta}) &= \frac{\partial}{\partial \sigma^2_\alpha} \log \int_{\mathcal{R}} f(\mathbf{w}) d\mu = \frac{\frac{1}{2}\int_{\mathcal{R}} \left[(\mathbf{w} - \mathbf{X}\boldsymbol{\beta})'\,\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{w} - \mathbf{X}\boldsymbol{\beta}) - \mathrm{tr}\left(\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\right)\right] f(\mathbf{w}) d\mu}{\int_{\mathcal{R}} f(\mathbf{w}) d\mu} \\ &= \frac{1}{2}\left[\mathrm{E}\left\{(\mathbf{w} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{w} - \mathbf{X}\boldsymbol{\beta})|\mathbf{y}\right\} - \mathrm{tr}\left(\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\right)\right], \end{aligned} \tag{2.3}$$

where tr($\mathbf{A}$) denotes the trace of a square matrix $\mathbf{A}$. It can be shown that the conditional expectation in (2.3) is equal to

$$\begin{aligned} &\mathrm{E}\left\{(\mathbf{w} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{w} - \mathbf{X}\boldsymbol{\beta})\middle|\mathbf{y}\right\} \\ &= \mathrm{E}\left(\mathbf{w}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{w}\middle|\mathbf{y}\right) - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\mathbf{V}^{-1}\mathrm{E}(\mathbf{w}|\mathbf{y}) + \boldsymbol{\beta}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta}, \end{aligned}$$

where $\mathrm{E}(\mathbf{w}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{w}|\mathbf{y}) = \sum_{i=1}^{n} \mathrm{E}(\mathbf{w}_i'\mathbf{V}_i^{-1}\mathbf{J}_i\mathbf{V}_i^{-1}\mathbf{w}_i|\mathbf{y})$, since $\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\mathbf{V}^{-1}$ is a block diagonal matrix with the i^{th} diagonal matrix $\mathbf{V}_i^{-1}\mathbf{J}_i\mathbf{V}_i^{-1}$ and $\mathbf{V}_i^{-1} = \mathbf{I} - \sigma^2_\alpha/(1 + T_i\sigma^2_\alpha)\mathbf{J}_i$. Note that

$$\mathrm{E}\left(\mathbf{w}_i'\mathbf{V}_i^{-1}\mathbf{J}_i\mathbf{V}_i^{-1}\mathbf{w}_i\middle|\mathbf{y}\right) = \mathrm{tr}\left(\mathbf{V}_i^{-1}\mathbf{J}_i\mathbf{V}_i^{-1}\mathrm{E}\left(\mathbf{w}_i'\mathbf{w}_i\middle|\mathbf{y}_i\right)\right).$$

Thus, the likelihood equations are the function of conditional expectations which are related to up to the second order moments of a multivariate truncated normal random vector. One may use the EM algorithm given in McCulloch (1994) or that of Chan and Kuk (1997) for another version of the EM

algorithm. Even though we may use one of these algorithms, the computation of the second order moments is essential.

To avoid the calculation of moments, many researchers approximate the conditional moments at this stage, and then apply the EM or variations of EM algorithms. This kind of approach includes the SEM (Celeux and Diebolt, 1985), the SAEM (Celeux *et al.*, 1996) and the MCEM (Wei and Tanner, 1990). However, the second order moments required from the EM can be calculated by the algorithm given in Manjunath and Wilhelm (2009), and it is implemented in a R package **tmvtnorm** (Wilhelm, 2015). Thus, such kind of approximation is not necessary in the probit–normal model. Nonetheless, it would be desirable to use the Newton-Raphson method rather than the EM algorithm, because the EM algorithm was slow and the Hessian matrix is required to compute the asymptotic standard error of the ML estimate; however, up to the fourth order moments are required to apply the Newton-Raphson method.

There is a history of the moment calculation for a multivariate truncated normal random vector. Using the moment generating function or recurrence relationships, many moment calculation methods have been proposed under various conditions, see Arismendi (2013). Among them, Kan and Robotti (2017) provide a suitable method for our needs. That is, using their algorithm, we can compute $\mathrm{E}(X_1^{k_1} \cdots X_n^{k_n} | a_i < X_i < b_i, i = 1, \ldots, n)$ for nonnegative integer values k_i satisfying $\sum_{i=1}^n k_i \leq 4$, where $\mathbf{X} = (X_1, \ldots, X_n)$ follows a multivariate normal distribution with arbitrary mean and positive definite covariance matrix. In what follows, we assume safely that necessary moments of truncated variables could be obtainable; however, their algorithm should compute $\sum_{i=1}^n 5^{T_i}$ conditional expectations for the panel probit model, which may be huge in some applications.

The Hessian matrix $\mathbf{J}(\boldsymbol{\theta})$ composed of the second order derivatives is

$$\mathbf{J}(\boldsymbol{\theta}) = \begin{pmatrix} \dfrac{\partial^2 \log L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}, & \dfrac{\partial^2 \log L}{\partial \boldsymbol{\beta} \partial \sigma_\alpha^2} \\ \dfrac{\partial^2 \log L}{\partial \sigma_\alpha^2 \partial \boldsymbol{\beta}}, & \dfrac{\partial^2 \log L}{\partial \sigma_\alpha^2 \partial \sigma_\alpha^2} \end{pmatrix},$$

where

$$\begin{aligned} \frac{\partial^2 \log L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} &= -\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} + \mathbf{X}'\mathbf{V}^{-1}\mathrm{Var}(\mathbf{w}|\mathbf{y})\mathbf{V}^{-1}\mathbf{X}, \\ \frac{\partial^2 \log L}{\partial \boldsymbol{\beta} \partial \sigma_\alpha^2} &= -\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\mathbf{V}^{-1}\mathrm{E}(\mathbf{w}|\mathbf{y}) + \frac{1}{2}\mathbf{X}'\mathbf{V}^{-1}\mathrm{Cov}\left(\mathbf{w}, \mathbf{w}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{w}|\mathbf{y}\right) \\ &\quad - \mathbf{X}'\mathbf{V}^{-1}\mathrm{Var}(\mathbf{w}|\mathbf{y})\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta}, \end{aligned}$$

and

$$\begin{aligned} \frac{\partial^2 \log L}{\partial \sigma_\alpha^2 \partial \sigma_\alpha^2} &= -\mathrm{E}\left(\mathbf{w}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\mathbf{V}^{-1}|\mathbf{y}\right) + \frac{1}{4}\mathrm{Var}\left(\mathbf{w}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{w}|\mathbf{y}\right) + \frac{1}{2}\mathrm{tr}\left(\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\right) \\ &\quad + 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\mathbf{V}^{-1}\left[\mathbf{Z}\mathbf{Z}'\mathbf{V}^{-1}\mathrm{E}(\mathbf{w}|\mathbf{y}) - \frac{1}{2}\mathrm{Cov}\left(\mathbf{w}, \mathbf{w}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{w}|\mathbf{y}\right)\right] \\ &\quad + \boldsymbol{\beta}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\mathbf{V}^{-1}\left[\mathrm{Var}(\mathbf{w}|\mathbf{y}) - \mathbf{V}\right]\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta}. \end{aligned}$$

To compute $\mathrm{Var}(\mathbf{w}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{w}|\mathbf{y})$ and $\mathrm{Cov}(\mathbf{w}, \mathbf{w}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{w}|\mathbf{y})$, it requires up to the fourth order moments. We do not talk about the detailed computation of these quantities, but once we can

Table 1: Estimates of UnionWage data

	Gauss–Hermite ($R = 64$)			Simulated ML (40 and 1600 Halton draws)						Theoretical ML		
	Estimate	Std.err	Score	Estimate	Std.err	Score	Estimate	Std.err	Score	Estimate	Std.err	Score
(Intercept)	−0.9210	0.2374	−0.2646	−0.9211	0.2381	−0.1730	−0.9210	0.2374	−0.0014	−0.9210	0.2374	2.29e−05
wage	0.5272	0.1363	−0.2995	0.5288	0.1367	−0.2892	0.5272	0.1363	−0.0030	0.5272	0.1363	1.15e−04
expr	−0.0409	0.0297	−2.2501	−0.0414	0.0297	−0.7111	−0.0409	0.0297	0.0026	−0.0409	0.0297	8.31e−06
rural	0.2333	0.1468	0.0378	0.2339	0.1472	−0.0082	0.2333	0.1468	−0.0015	0.2333	0.1468	3.27e−05
sigma	0.1395	0.3811	−3.4369	0.1151	0.2085	−1.2273	0.0983	0.2712	0.0284	0.0986	0.2695	−1.07e−04

compute $\mathbf{J}(\boldsymbol{\theta})$, then the ML estimate is obtained by the Newton-Raphson iterations, $\hat{\boldsymbol{\theta}}^{(m+1)} = \hat{\boldsymbol{\theta}}^{(m)} - \mathbf{J}^{-1}(\hat{\boldsymbol{\theta}}^{(m)})S(\hat{\boldsymbol{\theta}}^{(m)})$, $m = 0, 1, \ldots$, with an arbitral initial value $\hat{\boldsymbol{\theta}}_0$, where $S(\boldsymbol{\theta}) = (S_1'(\boldsymbol{\theta}), S_2(\boldsymbol{\theta}))'$.

Kennedy and Gentle (1980) and Searle *et al.* (2006) recommended to stop the iteration when Marquardt's criterion (Marquardt, 1963),

$$\left|\hat{\theta}_i^{(m+1)} - \hat{\theta}_i^{(m)}\right| \leq \epsilon_1 \left(\left|\hat{\theta}_i^{(m)}\right| + \epsilon_2\right), \quad \text{for all } i$$

is satisfied, where $\epsilon_1 = 10^{-4}$ and $\epsilon_2 = 10^{-3}$. Note that the ML estimate computed by a software is an approximation of the true ML estimate $\hat{\boldsymbol{\theta}}_T$, which satisfies $S(\hat{\boldsymbol{\theta}}_T) = \mathbf{0}$. Thus, the computed ML estimate $\hat{\boldsymbol{\theta}}$ should make $S(\hat{\boldsymbol{\theta}})$ close enough to $\mathbf{0}$. The large absolute value of score indicates that the estimate is not the ML estimate.

2.2. An example

To demonstrate the main feature of ML estimates from two differently approximated likelihood functions, we use Union Wage data in **pglm**, which consists of yearly observations on 12 variables of 545 individuals from 1980 to 1987, but to reduce the computational burden, we only use randomly selected 100 individual observations from 1980 to 1984. The binary dependent variable "union" is modelled by three variables "wage", "expr", and "rural." Refer to Croissant (2017) for detailed description of the variables.

Table 1 shows the ML estimates and their standard errors computed by two different approaches. We borrowed **pglm** and **Rchoice** packages to compute the ML estimates based on approximated likelihood functions, where the approximation is done by Gauss–Hermite quadratures or Monte Carlo integration. The ML estimate based on the theoretical likelihood function is also given in the table. For the last, we prepared a C++ program using RcppArmadillo (Eddelbuettel and Sanderson, 2014) under R environment (Eddelbuettel *et al.*, 2018).

pglm uses 20 quadrature points by default, but it could be small for some applications. Thus, we examined several numbers of quadrature point. It seems that the quadrature approximation is achieved with 20 points in this case. 20 and 64 quadrature points give almost same estimates. However, the number of draws in the simulated maximum likelihood approach have effected the ML estimate. By default, **Rchoice** uses 40 Halton draws, but it is apparently not sufficient. We have seen that the result of **Rchoice** gets closer to the theoretical ML estimate by increasing the number of draws. We have concluded that the approximation is achieved with 1600 draws after examining several number of draws.

The score values of theoretical ML are close to zero. Thus, we may consider the reported value as the computationally attainable ML estimate. Note that the two approximation approaches give similar values of estimate and standard error to ML for slope parameters, but give different estimates of variance component. It seems that the main difference between two approaches occurs in the estimation of the variance component. If we look at the value of the score calculated by (2.2) and

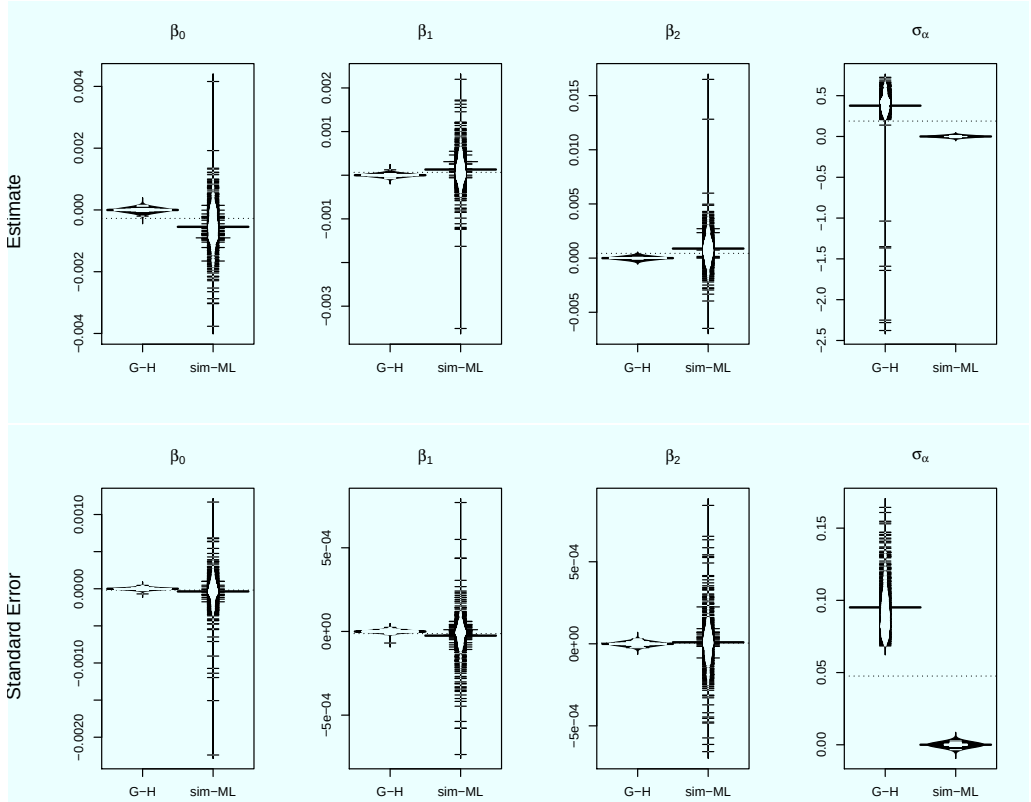

Figure 1: *Bean plots of differences in estimate and standard error.*

(2.3), **pglm** gives the estimate far from being the ML. It seems that **Rchoice** is better than **pglm** in terms of score, but it also seems that there is still room for improvement. However, the small change in the value of variance component may result in a different value of score (Table 1). It may be dangerous to use only the score value in a practical sense. If we look at the value of estimate itself, there is no difference in the slope parameter estimations. Both the Gauss–Hermite quadrature approximation and the simulated maximum likelihood method provide the same values of estimates and standard errors to the theoretical ML for the slope parameter, but the second approach seems better than the first for the estimation of the variance component.

3. Simulation study

To confirm the conjecture based on an example, we perform a simulation study under the design used in Harris *et al.* (2000) and other researchers. We generated the latent variables by

$$w_{it} = \beta_0 + \beta_1 x1_{it} + \beta_2 x2_i + \alpha_i + \epsilon_{it}, \quad i = 1, \ldots, n;\ t = 1, \ldots, T,\ \alpha_i \sim \text{NID}(0, 1),\ \epsilon_{it} \sim \text{NID}(0, 1).$$

The first regressor $x1_{it}$ is time-dependent to mimic a time series and was generated by an auto-regressive process, $x1_{it} = 0.1t + 0.5x1_{i,t-1} + u_{it}$, $u_{it} \sim U(-0.5, 0.5)$ with initial value generated by $x1_{i0} = 5 + 10u_{i0}$. The time invariant variable $x2_i$ was given to 0 or 1 whether $x2_i^*$ is less or greater than 0.5, where $x2_i^*$'s are uniform random numbers. The observed value was mapped from the latent

Table 2: Summary of simulation, $(\beta_0, \beta_1, \beta_2, \sigma_\alpha) = (1.5, -1, 1, 1)$

n	T	θ	Gauss–Hermite				Simulated ML				Maximum Likelihood			
			Ave. Est	MSE	Ave. Std	Cover.	Ave. Est	MSE	Ave. Std	Cover.	Ave. Est	MSE	Ave. Std	Cover.
30	5	β_0	1.5563	0.2121	0.4503	0.952	1.5564	0.2121	0.4504	0.952	1.5563	0.2121	0.4503	0.952
		β_1	−1.0579	0.0730	0.2457	0.954	−1.0578	0.0730	0.2457	0.954	−1.0578	0.0729	0.2457	0.954
		β_2	1.0396	0.3136	0.5040	0.940	1.0397	0.3139	0.5041	0.940	1.0395	0.3136	0.5040	0.940
		σ_α	1.3356	0.3715	0.4190	0.813	0.9621	0.0972	0.2961	0.986	0.9621	0.0972	0.2963	0.986
	10	β_0	1.5187	0.1461	0.3847	0.948	1.5188	0.1461	0.3847	0.948				
		β_1	−1.0148	0.0321	0.1769	0.940	−1.0148	0.0321	0.1769	0.940				
		β_2	1.0186	0.1926	0.4171	0.932	1.0185	0.1925	0.4172	0.932				
		σ_α	1.3554	0.2138	0.2891	0.714	0.9584	0.0454	0.2044	0.948				
50	5	β_0	1.5256	0.1212	0.3443	0.940	1.5251	0.1212	0.3443	0.938	1.5256	0.1212	0.3443	0.940
		β_1	−1.0250	0.0304	0.1823	0.972	−1.0248	0.0304	0.1822	0.974	−1.0249	0.0304	0.1823	0.972
		β_2	1.0565	0.1724	0.3863	0.936	1.0574	0.1725	0.3863	0.936	1.0565	0.1724	0.3863	0.936
		σ_α	1.3670	0.3463	0.3248	0.691	0.9896	0.0609	0.2297	0.980	0.9896	0.0609	0.2297	0.980
	10	β_0	1.5036	0.1027	0.3014	0.936	1.5037	0.1027	0.3014	0.936				
		β_1	−1.0075	0.0203	0.1433	0.960	−1.0074	0.0203	0.1433	0.960				
		β_2	1.0063	0.1118	0.3255	0.940	1.0061	0.1117	0.3255	0.940				
		σ_α	1.3805	0.1946	0.2244	0.516	0.9762	0.0255	0.1587	0.964				

variable by $y_{it} = \mathbf{1}[w_{it} \geq 0]$. The value of $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \sigma_\alpha)$ was set to $(1.5, -1, 1, 1)$. For each combination of $n = 30, 50$ and $T = 5, 10$, the ML estimates were computed 500 times; however, we were unable to compute the theoretical ML estimate when $T = 10$ because it took too much time. Thus, when $T = 10$, we abandon to find how well the numerical methods approximate the ML estimate, but try to find how well they estimate the true parameter in an average sense.

We computed the differences between the approximated estimate and the ML estimate during the simulation to see how well the Gauss–Hermite and simulated maximum likelihood methods approximate the ML estimate. The first row of Figure 1 is the bean plot of these quantities when $n = 50$ and $T = 5$. If the quantities are concentrated in zero, we may conclude that the approximation is well done.

It is observable the estimates of slope parameter given by the Gauss–Hermite method are almost same as the ML method. The Gauss–Hermite method provides better approximation than the simulated maximum likelihood method for slope parameter; however, the approximation of the latter method is also good in that the maximum bias is less than 0.02, which may be negligible in practice. Note also the simulated maximum likelihood method is better for the variance component. The Gauss–Hermite method has a tendency to give inflated estimate of variance component. It seems that this bias would be more critical. Thus, it may be concluded that the simulated maximum likelihood method is preferable to the Gauss–Hermit method in this study. The second row of Figure 1 is presented to see the approximation of the standard error of estimate. Both methods provide suitable standard error for the slope parameter, but the Gauss–Hermit method is unable to compute the proper standard error of variance component. It usually gives large value of estimate and standard error for the variance component. We also examined the case, $n = 30$ and $T = 5$, and reached the same conclusion since the bean plot of this case is essentially the same as Figure 1.

We wish to examine the behavior of two approaches, when T is moderately large, say $T = 10$. For this purpose, average of estimates $(1/500)\sum_{i=1}^{500}\hat{\theta}_i$, empirical mean square error $(1/500)\sum_{i=1}^{500}(\hat{\theta}_i - \theta)^2$, average of standard error $(1/500)\sum_{i=1}^{500}\mathrm{std}(\hat{\theta}_i)$ and the coverage probability of 95% Wald confidence interval $\hat{\theta} \pm 1.96\ \mathrm{std}(\hat{\theta})$ were computed for each parameter and presented in Table 2. The coverage probability is designed to see that the approach gives the proper estimate and standard error. If they are proper, then the coverage probability would be close to nominal level 95%. It is known that

the distribution of $\hat{\sigma}_\alpha$ is skewed to right, the log–transformed confidence interval was applied to the variance component for better approximation.

As before, two approaches behave very similarly in the estimation of slope parameters for all cases. In particular, they are essentially same as the ML, when $T = 5$. However, the Gauss–Hermite method do not give proper estimate and standard error of variance component in that the coverage probability is far from nominal level. It seems that this phenomenon will deteriorate when T is large. Unlike the Gauss–Hermite method, the simulated maximum likelihood method provides the proper estimate and standard error for all parameters and for all cases. This may justify that the simulated maximum likelihood method has an advantage over the Gauss–Hermite quadratures method.

4. Conclusion

The recently developed algorithm for the high order moments of a truncated multivariate normal distribution enables the exact computation of the ML estimate in various generalized linear mixed effects models such as the random-effects panel probit model; however, it is difficult to use the algorithm because of computational burden. Most statistical packages employ numerical integral methods, rather than the exact method, to approximate the ML estimate. Theoretically the approximation based on a numerical method can be improved to any precision, but it is also known that numerical integral approximation is inadequate for high dimensional integration in practice. As a result, different packages may give quite different outputs. Thus, we examine the adequacy of two most popular approximation methods, the Gauss–Hermite quadrature and the simulated maximum likelihood method, in the random-effects panel probit model. Based on a simulation study, we found that both methods have the ability of approximating the exact ML estimate and standard error quite well for regression coefficients when the time period is moderate. However, unlike the simulated maximum likelihood, the Gauss–Hermite quadrature does not adequately approximate both quantities for variance component. It has a tendency to give inflated estimates and standard errors of variance component. The Gauss–Hermite quadrature method shows slightly better performance than the other for the estimation of regression coefficient; however, we recommend the simulated maximum likelihood method with a sufficiently large number of Halton draws for practice.

Acknowledgement

This work was supported by Hanshin University research grant.

References

Arismendi JC (2013). Multivariate truncated moments, *Journal of Multivariate Analysis*, **117**, 41–75.

Bates D, Mächler M, Bolket BM, and Walker SC (2015). Fitting linear mixed–effects models using lme4, *Journal of Statistical Software*, **67**, 1–48.

Byrd RH, Lu P, Nocedal J, and Zhu C (1995). A limited memory algorithm for bound constrained optimization, *SIAM Journal on Scientific Computing*, **16**, 1190–1208.

Celeux G and Diebolt J (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem, *Computational Statistics Quarterly*, **2**, 73–82.

Celeux G, Chauveau D, and Diebolt J (1996). Stochastic versions of the EM algorithm: an experimental study in the mixture case, *Journal of Statistical Computation and Simulation*, **55**, 287–314.

Chan JSK and Kuk AYC (1997). Maximum likelihood estimation for probit–linear mixed models with correlated random effects, *Biometrics*, **53**, 86–97.

Croissant Y (2017). Package pglm: Panel generalized linear models, R package version 0.2-1. Available from: https://cran.r-project.org/web/packages/pglm/pglm.pdf

Eddelbuettel D, Francois R, Allaire J, Ushey K, Kou Q, Russell N, Bates D, and Chambers J (2018). *Seamless R and C++ Integration*. Available from: http://dirk.eddelbuettel.com/code/rcpp.html

Eddelbuettel D and Sanderson C (2014). RcppArmadillo: accelerating R with high-performance C++ linear algebra, *Computational Statistics and Data Analysis*, **71**, 1054–1063.

Halton JH (1964). Radical-inverse quasi-random point sequence, *Communications of the ACM*, **7**, 701–702.

Harris MN, Macquarie LM, and Siouclis AJ (2000). A comparison of alternative estimators for binary panel probit models, *Melbourne Institute Working Paper,* No 3. ISSN 1328-4991.

Hotelling H (1936). Relations between two sets of variates, *Biometrika*, **28**, 321–377.

Kan R and Robotti C (2017). On Moments of Folded and Truncated Multivariate Normal Distributions, Unpublished manuscript. Available from: https://sites.google.com/site/cesarerobotti/kr_JCGS.pdf

Kennedy JWJ and Gentle JE (1980) *Statistical Computing*, Marcel Dekker, Inc.

Lancaster T (2000). The incidental parameter problem since 1948, *Journal of Econometrics*, **95**, 391–413.

Marquardt DW (1963). An algorithm for least squares estimation of nonlinear parameters, *Journal of the Society for Industrial and Applied Mathematics*, **11**, 431–441.

Manjunath BG and Wilhelm S (2009). Moments calculation for the double truncated multivariate normal density (Working Paper).Available from: http://ssrn.com/abstract=1472153

McCulloch CE (1994). Maximum likelihood variance components estimation for binary data, *Journal of the American Statistical Association*, **89**, 330–335.

McCulloch CE (1996). Fixed and random effects and best prediction. In *Proceedings of the Kansas State Conference on Applied Statistics in Agriculture.*

McFadden D and Ruud PA (1994). Estimation by simulation, *The Review of Econometrics and Statistics*, **76**, 591–608.

Pinheiro J, Bates D, DebRoy S, and Sarkar D (2015). nlme: Linear and nonlinear mixed effects Models. R package version 3.1-122. Available from: http://CRAN.R-project.org/package=nlme

Sarrias M (2016). Discrete choice models with random parameters in R: The Rchoice Package, *Journal of Statistical Software*, **74**, 1–31.

Searle SR, Casella G, and McCulloch CE (2006). *Variance Components*, John Wiley & Sons, New York.

Wei GCG and Tanner MA (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms, *Journal of the American Statistical Association*, **85**, 699–704.

Wilhelm S (2015). Package tmvtnorm: Truncated Multivariate Normal and Student t Distribution. Available from: https://cran.r-project.org/web/packages/tmvtnorm/tmvtnorm.pdf

Wolfinger R and O'Connell M (1993). Generalized linear mixed models: a pseudo-likelihood approach, *Journal of Statistical Computation and Simulation*, **48**, 233–243.

Zhang H, Lu N, Feng C, Thurston SW, Xia Y, Zhu L, and Tu XM (2011). On fitting generalized linear mixed-effects models for binary responses using different statistical packages, *Statistics in Medicine*, **30**, 2562–2572.

Received March 9, 2019; Revised April 11, 2019; Accepted April 12, 2019