응용통계연구 제 13권 2호, 2000년, pp. 541-562

일반화된 선형 혼합 모형(GENERALIZED LINEAR MIXED MODEL: GLMM)에 관한 최근의 연구 동향

이준영[1)]

요약문

일반화된 선형 혼합 모형(GLMM)은 자료가 계수의 형태로 나타나는 범주형 자료의 경우, 혹은 집락의 형태나 과산포된 비정규 자료, 또는 비선형 모형에 따르는 자료를 다루기 위한 모형 설정에 사용된다. 본 연구에서는 이에 대한 개요와 더불어, 이 모형의 적합을 위해 제시된 통계적 기법들중 의사가능도(quasi-likelihood: QL)를 이용한 추정 방법 및 Monte-Carlo 기법을 이용한 추정 방법들에 대해 조사하였다. 또한 GLMM에 대한 현재의 연구 방향 및 앞으로의 연구 가능 주제들에 대해서도 언급하였다.

주요용어: 선형 모형, 혼합 모형, 의사 가능도, Monte Carlo 모의 실험.

1. 서론

전통적으로 범주형 자료(categorical data)나 계수 자료(count data)는 다음과 같은 두 가지 방식으로 분석되어 왔다. 첫째로는, 변환된 자료를 이용하여 선형 모형(Linear Model: LM)을 적합시키는 방법이다. 이때 자료의 분산을 안정시키기(stabilize) 위하여 변환을 사용한다. 둘째로는, 일반화된 선형 모형(Generalized Linear Model: GLM)을 통해 자료를 분석하는 방법이다. GLM은 반응값들의 평균에 대한 비선형 함수를 모형화할 수 있고, 많은 경우에 자료의 분포에 대한 보다 직접적인 모형화가 가능하다는 측면에서, LM에 비해 더 진보된 모형이라고 할 수 있겠다. 특히 자료의 분포에 대한 직접적인 모형화를 할 수 있다는 점은, 범주형 자료나 계수 자료의 경우 실험자가 표본 계획을 통해 자료의 분포를 조절할 수 있다는 면에서 특별한 중요성을 지닌다. 연속형 자료의 경우 집락 자료(clustered data)나 상관된 자료(correlated data)들은 선형 혼합 모형(Linear Mixed Model: LMM)을 통해 분석될 수 있다. 그러나 GLM 방법으로는 이러한 자료들에 대한 분석이 불가능하다. GLM의 또 하나의 단점은 자료가 가정된 분포하에서 얻어져야 하는 변동보다 더 큰 변동을 나타내는 과산포(overdispersion)의 문제를 해결하는데 어려움이 있다는 점이다. 한편, LMM은 분할구(split-plot) 모형, 시계열 회귀 모형, 공간 변동 모형(spatial variation model)등 그 적용 범위가 광범위하나, 정규 오차 가정을 필요로 한다는 점에서 적용의 근본적인 한계점을 지닌다.

1) (136-701) 서울특별시 성북구 안암동 5가 126-1, 고려대학교 의과대학 예방의학교실 및 환경의학연구소, 연구강사 E-mail: jyleeuf@mail.korea.ac.kr

일반화된 선형 혼합 모형(Generalized Linear Mixed Model: GLMM)은 위 두가지의 모형, 즉 GLM과 LMM을 연결시킴으로서, 집락 자료나 상관 자료, 그리고 범주형/계수자료에 대한 과산포 문제를 다룰 수 있게 해 주었으며, GLM 및 LMM을 각각 그 특수한 형태로 포함시킨다 (이 준영, 1999 참조). 이에따라 최근들어 GLMM을 사용한 자료분석이 경제학(Langford, 1994), 교육학(Goldstein, 1991), 의·약학 및 보건학(Cnaan, Laird와 Slasor, 1997; Daniels와 Gatsonis, 1999; Gibbons, et al., 1994), 사회학(Murphy와 Wang, 1998, Nee, 1996; Sampson, Raudenbush와 Earls, 1997)등 각 분야에서 활발하게 응용되고 있다.

Breslow와 Clayton(1993) 및 Wolfinger와 O'Connell(1993)등에 의해 체계화되기 시작한 GLMM은, 그 시작을 항목-답변 모형(item-response model)을 제안한 Rasch(1961)로부터 찾을 수 있다. 따라서 이 모형은 Rasch 모형이라고도 불린다. Rasch는 개체 효과들(subject effects)을 하나의 고정된 장애 모수들(nuisance parameters)로 간주하고, 이 모수들의 충분 통계량에 근거한 조건부 분포를 생각하였다. 반면, 이 개체 효과들을 - Rasch는 반대하였던 - 변량 효과들(random effects)로 간주하여 분석한 연구들이 Bock과 Aitkin(1981), Stiratelli, Laird와 Ware(1984)등에 의해 시도되었다. GLMM이라는 용어는 Gilmour, Anderson과 Rae(1985)에 의해 처음 사용되었으며, 이와 관련된 연구들로 Anderson과 Aitkin(1985), Harville과 Mee(1984), Laird(1978), 그리고 Montgomery, Richards와 Braun(1986)등을 들 수 있다. GLMM에 대한 소개 논문으로는 Stroup과 Kachman(1994), Kachman과 Stroup(1994)등이 있다. GLMM은 또한 모수적 경험적 베이지안 분석(parametric empirical Bayesian analysis)과도 관련이 있다. 즉 GLMM은 변량 효과에 대해 정규 사전확률(normal prior)을 생각하고, 자료의 주변 분포로부터 얻어지는 최대 가능도 추정값(maximum likelihood estimate: MLE)을 이용하여 초모수들(hyperparameters)의 추정값을 얻는 방법이라고 볼 수 있다. GLMM과 모수적 경험적 베이지안 분석과의 관계는 Booth와 Hobert(1998)에 간략히 언급되어 있다.

본 연구에서는 GLMM의 모형적 접근 방법에 대한 개략적 소개와 더불어, 현재 이루어지고 있는 연구의 방향 및 GLMM의 장, 단점에 대해 논하고, 앞으로의 연구 가능 주제들(McCulloch, 1999)에 대해 언급하고자 한다.

2. 모형 설정

GLMM의 틀 안에서 변량 효과를 설명하기 위한 방법으로, 초창기의 연구로서, 다음과 같은 세가지의 시도가 있었다. 첫째로는, 계산상의 편의를 위해 특수한 형태의 모형, 예를 들면 베타-이항 모형(beta-binomial model), 또는 포아송-감마 모형(Poisson-gamma model)을 설정하는 방법이다. 두번째로, 반응값들의 결합 주변 분포를 이용하여, 변량 효과들에 대한 직접적인 기술을 피하는 방법으로, 이를 주변 가능도적 접근(marginal likelihood approach)이라 한다. 세째로는, 변량 효과를 다루기 위한 전혀 다른 시도로서, 조건부 가능도적 접근(conditional likelihood approach)이 있다. 이 방법은 변량 효과들을 장애 모수로 간

주하고, 이들의 충분 통계량에 근거한 조건부 분포를 통하여 변량 효과들을 제거한 뒤 모수 효과에 대한 추론을 시도하는 방법이다. 하지만 변량 효과를 포함하고 있는 일반적인 상황에서 위의 가능도들을 최대화하기란 쉽지 않으므로, 의사 가능도(Quasi-Likelihood: QL)에 기초한 최대 가능도 추정법, Monte-Carlo 모의 실험에 기초한 최대 가능도 추정법, 그리고 구적(quadrature)을 이용하여 가능도 함수를 직접 근사하는 방법등이 개발되었다. QL에 기초한 방법중, Breslow와 Clayton(1993)은 벌칙 의사 가능도(Penalized Quasi-Likelihood: PQL)를 이용하는 방법과, 주변 의사 가능도(Marginal Quasi-Likelihood: MQL)를 이용하는 방법을 제시하였다. 본 연구에서는 PQL에 근거하는 최대 가능도 추정법에 대한 개요를 알아보고, Monte-Carlo 모의 실험 및 구적에 기초한 추정법에 대해 알아보도록 하겠다.

먼저 모형을 설정하기로 하자. GLMM은 다음의 4가지 성분(components)으로 이루어진다: 첫째로, 변량 효과가 주어진 상태에서의 반응값들의 분포 (주로 자연 지수족 내의 분포를 가정한다). 둘째로, 모수 효과와 변량 효과를 내포하고 있는 선형 추정량(linear predictor). 세째, 반응값들의 조건부 평균과 선형 추정량을 연결해주는 연결 함수(link function), 그리고 마지막으로 변량 효과에 대한 분포이다 (일반적으로 정규 분포를 가정한다). 이 경우, 변량 효과가 주어진 상태에서의 반응값 y_i, $i = 1, 2, \cdots, n$에 대한 확률 밀도 함수는

$$f_{y_i|\boldsymbol{u}}(y_i|\boldsymbol{u}, \boldsymbol{\beta}, \phi_i) = exp\left[\frac{1}{a(\phi_i)}(y_i\theta_i - c(\theta_i)) + d(y_i, \phi_i)\right] \tag{2.1}$$

로 표현되며, 이때 $\boldsymbol{u}$는 변량 효과 벡터, ϕ_i는 알려진 가중값 (예를 들어 이항 표본의 크기), 그리고 θ_i는 정준 모수를 나타낸다. $\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$라 가정하고, $\mu_i = E[y_i|\boldsymbol{u}]$, 그리고 연결함수를 $g(\cdot)$로 표현 한다면, GLMM에 의해, 반응값들의 조건부 평균(μ_i)과 선형 추정량(η_i)은,

$$g(\mu_i) = \eta_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{u} \tag{2.2}$$

의 관계를 가지며, 이때 $\boldsymbol{x}_i$는 설명 변수의 열벡터 값들을, 그리고 $\boldsymbol{z}_i$는 변량 효과들의 계수로 이루어진 열벡터를 의미한다. 여기서 $\boldsymbol{\Sigma}$는 $\boldsymbol{u}$의 분산-공분산 행렬로, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\sigma}^2)$, 즉 알려지지 않은 분산 성분들(variance components)의 벡터 $\boldsymbol{\sigma}^2$에 의존한다. 사실, 반응값들의 조건부 확률 분포가 자연 지수족에 속할 필요는 없다. 예를 들어 Breslow와 Clayton(1993)은 의사 가능도(quasi-likelihood)에 근거한 방법을 사용하여, 보다 포괄적인 분포족에 대해 다루었다. 또한 변량 효과에 대한 정규성 가정도 꼭 필요한 것은 아니다. 예를 들면, 우리는 Lee와 Nelder(1996)의 계층구조 일반화된 선형 모형(hierachical generalized linear model: HGLM)이나 Aitkin과 Francis(1995)의 비모수 혼합 분포(nonparametric mixing distribution)의 확장을 GLMM에 적용할 수가 있다. 단지 정규성의 가정이 사용되는 주된 이유는 변량 효과들 사이의 복잡한 공분산 구조에 대한 설정이 훨씬 간편해지기 때문이다. 마지막으로 가중값 ϕ_i 역시 꼭 알려진 값일 필요는 없다. 단지 해석상의 편의를 위해 가정되었을 뿐이다.

GLMM은 범주형 자료의 분석에서 자료가 집락의 형태로 얻어지는 경우에 특히 유용하다. 예를 들어 y_{ij}가 i번째 집락내 j번째 관찰값을 나타낸다고 하자. ($i = 1, 2, \cdots, t$; $j =$

$1, 2, \cdots, n_i$; $\sum_{i=1}^{t} n_i = N$). 그러면 모형 (2.2)의 선형 추정량은

$$g(\mu_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{u}_i \tag{2.3}$$

로 표현되며, $\boldsymbol{u}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$에 따르고, $\boldsymbol{u}_i$와 $\boldsymbol{u}_j$는 $i \neq j$ 일때 서로 독립이며, 따라서 $\boldsymbol{\Sigma}_i$는 대각행렬이 된다.

2.1. 이항 분포의 예

반복 측정된 베르누이 시행의 예로서, 항목-답변 모형을 생각해 보자. 한 개체가 여러가지 항목에 대해 "예"와 "아니오"의 답변을 한다고 가정하고, 이때 y_{ij}가 위 (2.3)에서 본 바와 같이 i번째 개체의 j번째 항목에 대한 답변 여부를 나타낸다고 하자. 그리고 각 개체의 답변 효과는 u_i로 표현하자. 그러면 이에 따른 적절한 모형은

$$\begin{aligned} &y_{ij}|u_i \sim Bernoulli(\mu_i), \\ &logit(\mu_i) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + u_i, \\ &u_i \sim ind.\, N(0, \sigma^2) \end{aligned}$$

이 될 것이다. 이때 변량 효과 u_i는 모형에서 개체내 답변값(반응)들간의 음이 아닌 상관을 설명해 주는 역할을 한다.

보다 간단한 예로, n명의 개체에 대해 "예"와 "아니오"의 답변을 하는 두가지 항목을 조사한 이항 짝진 자료(binary matched pairs)를 생각해 보자. (예를 들면, 특정 정당에 대한 지지도 변화의 추이를 알아보기 위해 동일한 유권자들 n명을 대상으로 서로 다른 두 시점에서 지지 여부를 조사하여 얻어지는 자료.) 이때 우리는 i번째 개체에 대해 두개의 종속된 이항표본(two dependent binomial sample)인 (y_{i1}, y_{i2})의 짝진 반응값을 얻게 된다. 여기서 y_{i1}은 첫번째 조사 시점에서의 i번째 개체의 답변 여부, 그리고 y_{i2}는 두번째 조사 시점에서의 답변 여부를 나타낸다. 이제 i번째 개체가 j번째 조사에서 "예"라는 답변을 할 확률을 μ_{ij}라 하고, 로짓 연결함수를 고려하자. 만일 두 조사 시점간에 답변 확률은 이질적인 반면 각 개체의 답변 로짓은 일정하다고 가정하면 다음과 같은 모형을 설정할 수 있다.

$$logit(\mu_{i1}) = \alpha + u_i, \quad logit(\mu_{i2}) = \alpha + \beta + u_i$$

이때 u_i는 $E(u_i) = 0$인 i번째 개체의 변량 효과이다. 즉 이 모형은 각 개체가 두번째 조사에서 "예"라 답변할 odds가 첫번째 조사에서의 odds의 e^{β}배가 됨을 상정한 것이다. 이와 더불어 $u_i \sim ind.N(0, \sigma^2)$이고, y_{ij}는 u_i가 주어진 상태에서 조건부 독립이라고 가정하면, 위 모형은 $\boldsymbol{\beta}' = (\alpha, \beta)$, $\boldsymbol{x}'_{i1} = (1, 0)$, $\boldsymbol{x}'_{i2} = (1, 1)$, $\boldsymbol{z}'_{i1} = \boldsymbol{z}'_{i2} = 1$을 가지는 모형 (2.3)의 특수한 경우가 된다.

2.2. 선형 혼합 모형의 예

선형 혼합 모형의 경우, 모형은

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}$$

으로 표현되고, $\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma_u})$, $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon})$이며 서로 독립이다. 따라서

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta},\ \boldsymbol{Z}\boldsymbol{\Sigma_u}\boldsymbol{Z}' + \boldsymbol{\Sigma_\epsilon})$$

가 되며, 조건부 분포를 이용하면

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{u} &\sim N(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \boldsymbol{\Sigma_\epsilon}) \\ \boldsymbol{u} &\sim N(\boldsymbol{0}, \boldsymbol{\Sigma_u}) \end{aligned}$$

로 표현할 수 있다. 이는 바로 항등 행렬을 연결 함수로 이용한 GLMM의 모형 표현 방식과 일치한다. 즉 LMM은 GLMM의 특수한 경우라 하겠다.

3. 모형 적합

GLMM의 모형 적합에 대한 기술상의 문제는 그 계산상의 복잡성때문에 많은 관심을 받고 있다. 예를 들어 주변 가능도적 접근의 측면에서 보면, 관찰값들의 결합 가능도 대신 변량 효과에 대한 적분을 통해서 이들 변량 효과들을 제거한(integrated out) 주변 가능도 함수를 이용하여 모수 $\boldsymbol{\beta}$와 $\boldsymbol{\Sigma}$의 최대 가능도 추정값을 얻는 방법이라고 상기했었다. 다시 말하면, 모형 (2.2)의 경우 모수 $\boldsymbol{\beta}$와 $\boldsymbol{\Sigma}$의 최대 가능도 추정값 $\hat{\boldsymbol{\beta}}$과 $\hat{\boldsymbol{\Sigma}}$은 다음의 주변 가능도 함수

$$l(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\boldsymbol{y}) = \int \prod_{i=1}^{n} f(y_i|\boldsymbol{u}, \boldsymbol{\beta}, \phi_i) f(\boldsymbol{u}|\boldsymbol{\Sigma}) d\boldsymbol{u} \tag{3.1}$$

를 최대화함으로서 얻어진다. 한편, 집락간 변량 효과들이 서로 독립인 경우에, 모형 (2.3)에 대한 주변 가능도 함수는

$$l(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\boldsymbol{y}) = \prod_{i=1}^{t} \int \prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{u}_i, \boldsymbol{\beta}, \phi_{ij}) f(\boldsymbol{u}_i|\boldsymbol{\Sigma}) d\boldsymbol{u}_i \tag{3.2}$$

로 표현될 수 있다. 이때 적분의 차원(dimension)은 주어진 집락내에 몇개의 변량 효과가 있는가에 따라 결정되게 된다. 또한 적분값내 결합 밀도 함수의 곱의 갯수는 집락내에 관찰값들의 수가 몇개인가에 비례해서 증가하게 된다. 거의 모든 경우에, 이러한 적분을 실제로 시도하기란 불가능하고, 따라서 의사 가능도 함수를 이용하거나, 모의 실험 또는 수치적 근사 방법을 사용하여 적분의 근사값을 얻고 이를 다시 수치적으로 최대화시켜 – 예를 들면 Newton-Rapson 방법등을 사용하여 – 모수의 추정값을 얻게 된다. 이에대한 보다 상세한 내용을 알아보기로 하자.

3.1. 벌칙 의사가능도(PQL) 접근 방법

GLMM을 모형 적합하는 방법론에 대한 연구는 Breslow와 Clayton(1993) 및 Wolfinger와 O'Connell(1993)에 의해 본격적으로 실시되었다. 본 연구에서는 Breslow와 Clayton의 PQL을 이용한 GLMM 적합 방법에 대해 알아보도록 하겠다.

편의상, 연결 함수는 정준 연결(canonical link)을 생각하자. (정준 연결에 대한 자세한 내용은 McCullagh와 Nelder(1989)를 참조하기 바란다.) 따라서 $\eta_i = \theta_i$가 된다. 반응값들에 대한 결합 밀도 함수가 (2.1)로 표현될때, 주변 가능도 함수 (3.1)은

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\boldsymbol{y}) &\approx |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \int \exp^{-k(\boldsymbol{u})} d\boldsymbol{u}, \\ k(\boldsymbol{u}) &= \sum_{i=1}^{n} \frac{(y_i\eta_i - c(\eta_i))}{a(\phi_i)} + \frac{1}{2}\boldsymbol{u}'\boldsymbol{\Sigma}^{-1}\boldsymbol{u} \end{aligned}$$

로 표현된다. 식 (2.2)에서 $\frac{d\eta_i}{d\boldsymbol{u}} = \boldsymbol{z}_i$이므로

$$\begin{aligned} k'(\boldsymbol{u}) &= -\sum_{i=1}^{n} \frac{(y_i - c'(\eta_i))\boldsymbol{z}_i}{a(\phi_i)} + \boldsymbol{\Sigma}^{-1}\boldsymbol{u} \text{ 이고} \\ k''(\boldsymbol{u}) &= \sum_{i=1}^{n} \frac{c''(\eta_i)\boldsymbol{z}_i\boldsymbol{z}_i'}{a(\phi_i)} + \boldsymbol{\Sigma}^{-1} \\ &= \boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z} + \boldsymbol{\Sigma}^{-1} \end{aligned}$$

가 된다. 여기서 $\boldsymbol{Z}'$은 변량 효과들의 계획 행렬(design matrix)을 나타내고, $\boldsymbol{W}$는 i번째 대각 원소가 $\frac{c''(\eta_i)}{a(\phi_i)}$인 대각 행렬이다. 바로 이 대각 원소들이 GLMM의 적합에서 반복 가중된 최소 제곱 추정값(iteratively weighted least squares estimates: IWLS)을 얻기위해 사용되는 가중값들이다. Taylor 확장에 의해 $k(\boldsymbol{u})$는

$$k(\boldsymbol{u}) = k(\tilde{\boldsymbol{u}}) + (\boldsymbol{u} - \tilde{\boldsymbol{u}})'k'(\tilde{\boldsymbol{u}}) + \frac{1}{2}(\boldsymbol{u} - \tilde{\boldsymbol{u}})'k''(\tilde{\boldsymbol{u}})(\boldsymbol{u} - \tilde{\boldsymbol{u}}) + O(\|\boldsymbol{u} - \tilde{\boldsymbol{u}}\|)$$

로 표현되며, $k'(\tilde{\boldsymbol{u}}) = 0$인 $\tilde{\boldsymbol{u}}$을 생각할 때, 잔류항을 무시한다면, 주변 가능도 함수 (3.1)은

$$l(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\boldsymbol{y}) \propto |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp^{-k(\tilde{\boldsymbol{u}})} \int \exp^{-\frac{1}{2}(\boldsymbol{u}-\tilde{\boldsymbol{u}})'k''(\tilde{\boldsymbol{u}})(\boldsymbol{u}-\tilde{\boldsymbol{u}})} d\boldsymbol{u} \tag{3.3}$$

의 형태를 가진다. 여기서 적분내의 값은 $N(\tilde{\boldsymbol{u}}, [k''(\tilde{\boldsymbol{u}})]^{-1})$에 따르는 확률 변수의 확률 밀도 함수 핵(kernel)으로 간주될 수 있으므로, 자료의 실제 주변 로그 가능도 함수는

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\boldsymbol{y}) &= -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\log|k''(\tilde{\boldsymbol{u}})| - k(\tilde{\boldsymbol{u}}) \\ &= -\frac{1}{2}\log|\boldsymbol{I} + \boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z}\boldsymbol{\Sigma}| - \sum_{i=1}^{n} \frac{(y_i\eta_i - c(\eta_i))}{a(\phi_i)} - \frac{1}{2}\tilde{\boldsymbol{u}}'\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{u}} \end{aligned} \tag{3.4}$$

가 될 것이다. 이때 가중치 행렬 $\boldsymbol{W}$의 원소들이 작은 값을 취한다는 가정을 하면 이 주변 로그 가능도 함수의 첫번째 항은 무시될 수 있고, 따라서 다음의 벌칙 의사 가능도 (Penalized Quasi-Likelihood: PQL)가 얻어지게 된다.

$$-\sum_{i=1}^{n}\frac{(y_i\eta_i - c(\eta_i))}{a(\phi_i)} - \frac{1}{2}\boldsymbol{u}'\boldsymbol{\Sigma}^{-1}\boldsymbol{u}. \tag{3.5}$$

모수 $\boldsymbol{\beta}$를 추정하기 위해서 PQL (3.5)을 $\boldsymbol{\beta}$와 $\boldsymbol{u}$에 대해 편미분하면, 모수들에 대한 다음의 점수 함수(score function)가 얻어진다.

$$\sum_{i=1}^{n}\frac{(y_i - c'(\eta_i))\boldsymbol{x}_i'}{a(\phi_i)} = 0, \tag{3.6}$$

$$\sum_{i=1}^{n}\frac{(y_i - c'(\eta_i))\boldsymbol{z}_i'}{a(\phi_i)} = \boldsymbol{\Sigma}^{-1}\boldsymbol{u}. \tag{3.7}$$

여기서 식 (3.7)은 $\boldsymbol{u}$의 분산-공분산 행렬 $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\sigma}^2)$의 원소들인 분산 성분 벡터 $\boldsymbol{\sigma}^2$에 의존하고 있으므로, 이를 추정하기 위해, Breslow와 Clayton(1993)은 선형 혼합 모형의 분석 방법을 이용하여, 다음과 같은 추정 방정식의 사용을 제안하였다. 즉, Patterson과 Thompson(1971)이 구한 제한된 최대 가능도(restricted maximum likelihood: REML) 함수

$$-\frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}\log|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}| - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) \tag{3.8}$$

(여기서 $\boldsymbol{V} = \boldsymbol{W}^{-1} + \boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}'$이고, $\boldsymbol{X}$는 모형 (2.2)에서 $\boldsymbol{x}_i'$을 i번째 행으로 가지는 행렬이다)를 $\boldsymbol{\sigma}^2$으로 편미분하여 Harville(1977)이 구한 추정 방정식

$$-\frac{1}{2}[(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial \sigma_j^2}\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) - tr(\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial \sigma_j^2})] = 0 \tag{3.9}$$

을 사용한다. (여기서 $\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}$이다.) 이때 Fisher 정보 행렬은

$$F = \{\mathrm{F}_{jk}\},\ \mathrm{F}_{jk} = -\frac{1}{2}tr\left(\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial \sigma_j^2}\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial \sigma_k^2}\right) \tag{3.10}$$

이 된다. (이에대한 자세한 내용은 Searle, Casella와 McCulloch, 1992를 참조하기 바란다.)

즉 Breslow와 Clayton이 제안한 GLMM 적합을 위한 PQL 접근 방식은, 먼저 n개의 자료들이 서로 독립이라는 가정하에서 (즉, $\boldsymbol{\Sigma} = \boldsymbol{I}(\boldsymbol{\sigma}^2)$) 일반적인 GLM 방법을 이용하여 $\boldsymbol{\beta}$의 초기 추정값 $\hat{\boldsymbol{\beta}}^{(0)}$을 구하고, 자료로부터 이 $\hat{\boldsymbol{\beta}}^{(0)}$를 이용하여 잔차들(residuals)을 얻은 뒤, 잔차들로부터 $\boldsymbol{\Sigma}$의 원소들, $\boldsymbol{\sigma}^2$에 대한 초기 추정값 $\hat{\boldsymbol{\sigma}}^{2(0)}$를 구한다. 이렇게 얻어진 모수들의 초기값들을 사용하여, 앞의 추정 방정식들 (3.6)과 (3.7)로부터, 새로운 추정값 $(\hat{\boldsymbol{\beta}}^{(1)}, \hat{\boldsymbol{u}}^{(1)})$을 구한 다음, 이들로부터 식 (3.9)과 (3.10)을 이용하여 $\boldsymbol{\sigma}^2$의 새로운 추정값 $\hat{\boldsymbol{\sigma}}^{2(1)}$을 구한다. 이렇게 얻어진 $\hat{\boldsymbol{\sigma}}^{2(1)}$ 값들을 다시 (3.6)과 (3.7)에서 이용하는 반복을 이들 추정값들이 수렴할 때까지 실시하여 원하는 모수들 $\boldsymbol{\beta}$와 $\boldsymbol{\Sigma}$에 대한 최대 가능도 추정값을 얻는 것이다.

3.2. 최근에 개발되는 새로운 추정 방법들

PQL을 이용한 추정 방법은 계산이 상대적으로 쉽고 변량 효과의 분산이 상대적으로 작을때 - 즉, 모형이 모수효과에 의해 지배(dominate)될 때 - 에는 잘 적용되지만, 이항 짝진 자료같은 희박 자료(sparse data)의 경우에는 그다지 만족할만한 결과를 제공하지 못한다 (Breslow와 Lin, 1995; McCulloch, 1997). 이에대한 대안으로 McCulloch(1997)는 다른 방법들 - MCEM 알고리즘(Monte-Carlo EM algorithm) 방법, MCNR(Monte-Carlo Newton-Rapson) 방법, SML(Simulated Maximum Likelihood) 방법등 - 을 제안하였다. 이들에 대해 알아 보자.

MCEM 알고리즘 방법은 변량 효과들을 결측치로 간주한 뒤, $\boldsymbol{y}$의 기대값을 구하는 단계에서 Monte-Carlo 근사를 사용하여 EM 알고리즘을 적용하는 방법이다. 이 방법은 Wei와 Tanner(1990)가 제안하였으며 그 내용은 다음과 같다. 자료가 주어졌을때 변량 효과에 대한 조건부 밀도 함수를 $f_{\boldsymbol{u}|\boldsymbol{y}}(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{\Sigma})$로 표현하자. 또 $f_{\boldsymbol{y},\boldsymbol{u}}(\boldsymbol{y},\boldsymbol{u}|\boldsymbol{\beta},\boldsymbol{\Sigma})$를 $\boldsymbol{y}$와 $\boldsymbol{u}$의 결합 밀도 함수라 하고, 이때 $\boldsymbol{y}$의 주변 가능도 함수는 (3.1)로 표현된다고 하자. 그러면 MCEM 알고리즘을 이용한 모수들의 최대 가능도 추정값들은 다음의 계산 절차에 의해서 얻어진다.

1. 초기값 $\boldsymbol{\beta}^{(0)}$와 $\boldsymbol{\Sigma}^{(0)}$를 구한다.
2. $f_{\boldsymbol{u}|\boldsymbol{y}}(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{\beta}^{(m)},\boldsymbol{\Sigma}^{(m)})$로부터 $\boldsymbol{u}^{(1)},\boldsymbol{u}^{(2)},\cdots,\boldsymbol{u}^{(s)}$를 발생시킨다.
3. $\frac{1}{s}\sum_{i=1}^{s}\log f_{\boldsymbol{u}|\boldsymbol{y}}(\boldsymbol{u}^{(i)}|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{\Sigma}) \approx E[\log f_{\boldsymbol{u}|\boldsymbol{y}}(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{\Sigma})|\boldsymbol{y},\boldsymbol{\beta}^{(m)},\boldsymbol{\Sigma}^{(m)}]$을 최대화하는 값 $\boldsymbol{\beta}^{(m+1)}$과 $\boldsymbol{\Sigma}^{(m+1)}$을 구한다.
4. $m=m+1$이라 놓는다.
5. 단계 $2\sim 4$를 $\boldsymbol{\beta}^{(i)}$와 $\boldsymbol{\Sigma}^{(i)}$가 수렴할 때까지 반복해서 얻어지는 $\boldsymbol{\beta}^{(m+1)}$과 $\boldsymbol{\Sigma}^{(m+1)}$이 구하는 추정값이 된다.

한편 $f_{\boldsymbol{u}|\boldsymbol{y}} \propto f_{\boldsymbol{y}|\boldsymbol{u}}\cdot f_{\boldsymbol{u}}$라는 점을 이용하면, 위의 단계 3은 아래의 3a와 3b로 대체될 수 있다.

3a. $\frac{1}{s}\sum_{i=1}^{s}\log f_{\boldsymbol{y}|\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{u}^{(i)},\boldsymbol{\beta})$를 최대화하는 값 $\boldsymbol{\beta}^{(m+1)}$을 구한다.

3b. $\frac{1}{s}\sum_{i=1}^{s}\log f_{\boldsymbol{u}}(\boldsymbol{u}^{(i)}|\boldsymbol{\Sigma})$를 최대화하는 값 $\boldsymbol{\Sigma}^{(m+1)}$을 구한다.

여기서 주의해야 할 단계는 2번 단계이다. 일반적으로 $f_{\boldsymbol{u}|\boldsymbol{y}}$로부터 독립 표본들을 생성시키기란 쉽지 않다. 이에 대한 대안으로 Markov Chain에서 얻어지는 종속 표본들을 사용하는 방법이 있다 (Markov Chain Monte-Carlo: MCMC). 이들 종속 표본들을 얻기 위해 McCulloch(1994), Chan과 Kuk(1997)등은 Gibbs sampler를 사용하는 방법을, McCulloch(1997)는 $f_{\boldsymbol{u}|\boldsymbol{y}}$에 대한 밀도 함수로 $f_{\boldsymbol{u}}(\boldsymbol{u}|\boldsymbol{\Sigma})$와 더불어 Metropolis-Hasting 알고리즘을 사용하는 방법을 제안하였다. 한편, Booth와 Hobert(1999)는 기각 표집(rejection sampling)과 중요부

표집(importance sampling)을 이용한 독립 표본들을 사용하였다. 이들은 종속 표본들을 사용하는 대신에 독립 표본들을 사용함으로서 얻어지는 장점으로 E-step인 3번 단계에서 Monte-Carlo 오차를 측정할 수 있게 된다고 주장하였다. 또한 E-step에서 정확한 값을 계산하기 위하여 s를 얼마나 크게 잡아야 하는지에 대해서 언급하고, 이 방법이 프로그램내에서 Monte-Carlo 표본의 갯수와 반복의 갯수가 자체적으로 결정되는 "완전 자동화(fully automated)"된 특성을 가진다는 점도 역설하였다.

MCNR 방법은 주변 가능도 함수 (3.1)을 최대화하는 방법이 다음의 두 방정식을 푸는 것과 동일하다는 점을 이용한 것이다.

$$\begin{aligned} E\left[\frac{\partial}{\partial \boldsymbol{\beta}} \log f_{\boldsymbol{y}|\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{u},\boldsymbol{\beta}) \,\middle|\, \boldsymbol{y}\right] &= 0 \\ E\left[\frac{\partial}{\partial \boldsymbol{\Sigma}} \log f_{\boldsymbol{u}}(\boldsymbol{u}|\boldsymbol{\Sigma}) \,\middle|\, \boldsymbol{y}\right] &= 0. \end{aligned}$$

일반적으로 이 두 방정식은 각각 독립적으로 풀며, 처음것은 적분값의 Monte-Carlo 추정값을 이용한 Newton-Raphson 방식을 통하여, 그리고 두번째 것은 해석적으로(analytically) 계산하게 된다.

SML 방법은 주변 가능도 함수 (3.1)을 중요부 표집 알고리즘을 이용하여 계산하는 방법이다. 즉 모의 표본이 얻어질 대상 밀도 함수를 $h(\boldsymbol{u})$, 얻어진 모의 표본값을 $\boldsymbol{u}^{(i)}$라 할 때, 가능도 함수는

$$\begin{aligned} l(\boldsymbol{\beta},\boldsymbol{\sigma}|\boldsymbol{y}) &= \int f_{\boldsymbol{y}|\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{u},\boldsymbol{\beta}) f_{\boldsymbol{u}}(\boldsymbol{u}|\boldsymbol{\sigma}) d\boldsymbol{u} \\ &= \int \frac{f_{\boldsymbol{y}|\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{u},\boldsymbol{\beta}) f_{\boldsymbol{u}}(\boldsymbol{u}|\boldsymbol{\sigma})}{h_{\boldsymbol{u}}(\boldsymbol{u})} h_{\boldsymbol{u}}(\boldsymbol{u}) d\boldsymbol{u} \\ &\approx \sum_{i=1}^{n} \frac{f_{\boldsymbol{y}|\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{u}^{(i)},\boldsymbol{\beta}) f_{\boldsymbol{u}}(\boldsymbol{u}^{(i)}|\boldsymbol{\sigma})}{h_{\boldsymbol{u}}(\boldsymbol{u}^{(i)})} \end{aligned}$$

로 근사된다. 따라서 이를 수치적으로 최대화하면 모수들의 최대 가능도 추정값을 얻을 수 있게 된다. McCulloch(1997)는 만일 주변 가능도 함수 (3.1)이 알려지지 않은 다른 모수를 포함하고 있다면, 가능도의 근사 단계와 그 다음 단계인 모수를 추정하는 단계, 이 둘을 번갈아 가면서 반복함으로서 최대 가능도 추정값들을 얻을 수 있다고 하였다.

마지막으로, GLMM을 적합시키는 방법중 하나인, Gauss-Hermite 구적(quadrature)을 사용하는 방법에 대해 알아 보도록 하자. 구적 방법은 적분을 합으로 근사시키는 방법이다. 특별히 Gauss-Hermite 구적 방법은 $\int_{-\infty}^{\infty} h(x)\exp(-\frac{x^2}{2\sigma^2})dx$ 형태의 적분값을 합으로 근사시키는 방법이다(Liu와 Pierce, 1994 참조). 주변 가능도 함수 (3.1)이 하나의 변량 효과나 또는 여러 개의 독립적인 변량 효과들을 가지게 된다면 바로 위와 같은 형태를 가지게 됨을 주목하자. 이때 근사는 $\int_{-\infty}^{\infty} h(x)\exp(-\frac{x^2}{2\sigma^2})dx \approx \sum_{i=1}^{m} w_i h(\sqrt{2\sigma}x_i)$의 형태로 되며, 여기서 마디(node)들 x_i는 m차 Hermite 다항식의 0의 값들이고, w_i는 이에 해당하는 가중값들이다.

가중값과 마디값들은 표에서 얻거나, 되풀이 계산(recursive calculation)을 통해 얻을 수 있다. 이 근사는 $h(x)$가 m차이거나 또는 그보다 낮은 차원의 다항식일 경우 정확한 근사값을 제공한다. 그러므로 마디의 갯수 m은 $h(x)$가 Taylor 전개에 의해 적절하게 근사되는 경우의 항(term)들의 갯수 근처에서 결정하면 된다. 주변 가능도 함수 (3.1)의 근사값이 얻어지면, 최대 가능도 추정값은 Newton-Rapson이나 Fisher의 점수화 방법(scoring method)등 최대화 프로그램을 사용하여 얻을 수 있다. 물론 Gauss-Hermite 구적 방법은 여러가지 수치적분 방법중 하나일 뿐이며, 단지 GLMM의 틀 안에서 볼때, 이 방법이 변량 효과의 정규성 가정이라는 점을 이용할 수 있기 때문에 고려된 것이다. SAS의 PROC MIXED를 사용하여 GLMM을 적합시키는 GLIMMIX Macro는, Wolfinger와 O'Cornell의 추정 방법을 이용하여 얻어지는 가능도 함수를 근사시키며, 이를 위해 일반화된 추정 방정식(Generalized Estimating Equation: GEE)을 사용한다 (Littell, et al, 1996). 반면에 최근 SAS의 제 7 버젼에서 소개된 PROC NLMIXED는 조정된 Gauss-Hermite 구적(adaptive Gauss-Hermite quadrature)방법을 적용하여 직접 가능도 함수를 근사시키는 방법을 사용한다 (Pinheiro와 Bates, 1995; SAS, 1999).

이외에도 변량 효과를 가지는 비례-오즈 모형(proportional odds model)에 사용될 수 있는 Hedeker와 Gibbons(1994)의 MIXOR (www.uic.edu/~hedeker/mix.html) FORTRAN 프로그램, 다수준 모형(multilevel model) 분석에 사용될 수 있는 Prosser, Rasbash와 Goldstein(1995)의 MLn (www.ioe.ac.uk/multilevel), Gibbs sampling을 사용하여 MCMC 방법을 통한 Bayesian 분석을 할 수 있는 BUGS (www.mrc-bsu.cam.ac.uk/bugs), PQL을 사용하는 HLM (Scientific Software International, Chicago), 집락 효과항을 제거하기 위해 조건부 ML 방법으로 접근하는 LogXact (Cytel Software, Cambridge, MA), 변량 효과 로지스틱 모형에 사용될 수 있는 EGRET등이 있다. 참고로 다수준 모형에 관한 소프트웨어 비교는 Zhou, Perkins와 Hui(1999)를 참조하기 바란다.

3.3. 요약

이상에서 볼 수 있는 바와 같이, 모형을 적합시키는 각 방법들은 시간상의 소모와 계산상의 복잡함을 내포한다. Breslow와 Clayton(1993)이나 Wolfinger와 O'Connell(1993)의 근사 방법들은 선형 혼합 모형을 사용하는 편리함과 SAS의 GLIMMIX Macro를 이용할 수 있다는 장점때문에 많이 애용된다. 하지만 이 방법들은 불일치 추정값을 제공하거나, 분산성분이 클때 근사가 적절하게 이루어지지 못한다는 단점을 가진다 (Breslow와 Lin, 1995; Jiang, 1998). 수치 적분 방법은 구적점들(m)을 증감시키면서 오차의 크기를 조절할 수 있는 장점이 있으나, 고차원 적분에는 정확한 결과를 얻을 수 없다는 단점을 지닌다. 한편 두 접근 방법들 모두 적분값들의 근사를 하는 단계에서 오차에 대한 정보를 얻을 수 없다는 면에 있어서, Booth와 Hobert(1999)의 방법과 비교된다.

최대 가능도 추정값을 얻은뒤 모수들에 대한 추론을 하기 위해서 그들의 점근적 성질

들(asymptotic properties)을 이용할 수 있다. 모수들의 표준 오차는 관찰된 정보 행렬이나 Hessian 행렬로부터 얻을 수 있다. 이 값들은 일반적으로 최대 가능도 추정값을 얻을 때 사용된 최대화 알고리즘내에서 얻어진다. 또한 추정값들의 점근적 정규성을 이용하여 모수들에 대한 추론도 할 수 있다.

모수들에 대한 해석 범위의 문제 또한 관심을 가져야 할 사항이다(이 준영, 1999 참조). 모형 (2.3)을 보자. 모수 $\boldsymbol{\beta}$는 변량 효과 $\boldsymbol{u}_i$가 주어졌을 때 반응값들의 조건부 분포에 의해 명시될 수 있다. 이때 $\hat{\boldsymbol{\beta}}$은 변량 효과들의 해당 수준값들에 대한 모수 효과의 추정값이다. 즉, 만일 변량 효과들이 얻어진 실험단위가 개체(subject)라면, $\hat{\boldsymbol{\beta}}$은 개체 중심적(subject specific: SS) 효과를 의미한다. 일반적으로 이 값들은 주변 분포에 의해서 얻어지는 효과와는 일치하지 않는다. 왜냐하면 모든 가능한 전체 변량 효과의 평균에 근거한 관찰값들의 주변 평균값은

$$E(y_{ij}) = E[E(y_{ij}|\boldsymbol{u}_i)] = E[g^{-1}(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{u}_i)]$$

이 되고, 이는 연결 함수 g가 항등 연결(identity link)이거나, 변량 효과가 퇴화(degenerate)하지 않는 이상 $E[g^{-1}(\boldsymbol{x}'_{ij}\boldsymbol{\beta})]$와 같지 않기 때문이다. 이 경우 주변 분포에 의해서 얻어진 $\boldsymbol{\beta}$의 추정값은 집단 평균적(population averaged: PA) 효과를 나타낸다. 따라서 Breslow와 Clayton(1993)은 집단 평균적 효과를 얻는 것이 주 관심사인 경우에는 변량 효과의 조건부 분포에 근거해서 GLMM을 적합시키기 보다는 변량 효과의 주변 분포에 근거해서 적합하는 방식이 합리적이라고 주장한다. Neuhaus, Kalbfleisch와 Hauck(1994)및 Zeger, Liang과 Albert(1988)는 개체 중심적 효과와 집단 평균적 효과간의 근사적 관계에 대해 언급하였다.

4. 예측

관찰되지 않는 변량 효과의 추정값을 얻기 위해서는 반응값에 대한 예측값을 이용해야만 한다. 변량 효과의 추정에 관한 많은 논점들이 Robinson(1991)에 나타나 있다. 변량 효과에 대한 추정값은

$$E(\boldsymbol{u}|\boldsymbol{y}) = \frac{\int \boldsymbol{u} \prod_{i=1}^{n} f(y_i|\boldsymbol{u}, \hat{\boldsymbol{\beta}}, \phi_i) f(\boldsymbol{u}|\hat{\boldsymbol{\Sigma}}) d\boldsymbol{u}}{\int \prod_{i=1}^{n} f(y_i|\boldsymbol{u}, \hat{\boldsymbol{\beta}}, \phi_i) f(\boldsymbol{u}|\hat{\boldsymbol{\Sigma}}) d\boldsymbol{u}}$$

로 표현된다. 마찬가지로 집락간 변량 효과들이 서로 독립인 경우에 i번째 집락의 변량 효과에 대한 추정값은

$$E(\boldsymbol{u}_i|y_{i1}, y_{i2}, \cdots, y_{in_i}) = \frac{\int \boldsymbol{u}_i \prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{u}_i, \hat{\boldsymbol{\beta}}, \phi_{ij}) f(\boldsymbol{u}_i|\hat{\boldsymbol{\Sigma}}) d\boldsymbol{u}_i}{\int \prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{u}_i, \hat{\boldsymbol{\beta}}, \phi_{ij}) f(\boldsymbol{u}_i|\hat{\boldsymbol{\Sigma}}) d\boldsymbol{u}_i}$$

가 된다. 이때 우리는 모수 $\boldsymbol{\beta}$와 $\boldsymbol{\Sigma}$에 대한 정보를 모르기때문에, 그 추정값 $\hat{\boldsymbol{\beta}}$과 $\hat{\boldsymbol{\Sigma}}$을 사용하게 될 것이다. 그러나 이 경우 반응값의 예측값에 대한 표준 오차를 얻을 수가 없기때문에,

이 표준 오차의 추정값으로 분산 $Var(\boldsymbol{u}_i|y_{i1}, y_{i2}, \cdots, y_{in_i}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ 대신, 모수 $\boldsymbol{\beta}$와 $\boldsymbol{\Sigma}$가 알려져 있다는 가정하에서 $\hat{\boldsymbol{\beta}}$과 $\hat{\boldsymbol{\Sigma}}$를 이용한 $Var(\boldsymbol{u}_i|y_{i1}, y_{i2}, \cdots, y_{in_i}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}})$를 사용해 왔다. Booth와 Hobert(1998)는 이러한 "순수"("naive") 추정값이 $\hat{\boldsymbol{\beta}}$과 $\hat{\boldsymbol{\Sigma}}$의 표집 변동을 설명해 주지 못하기 때문에 실제 분산을 과소추정(underestimate)하게 된다고 보았다. 이에 대한 대안으로 그들은 "예측의 조건부 평균 제곱 오차(Conditional Mean Square Error of Prediction: CMSEP)"를 예측값의 분산으로 사용할 것을 제안하였다. 예를 들어 y_i가 n_i개의 표본에서 실제 비율 π_i를 가지고 나타나는 계수 자료라고 하면, 이에 대한 적절한 모형은

$$y_i \sim B(n_i, \pi_i),\ logit(\pi_i) = \alpha + u_i,\ i = 1, 2, \cdots, t \tag{4.1}$$

이 될 것이며, 이때 $u_i \sim N(0, \sigma^2)$이고, u_i와 $u_j (i \neq j)$는 서로 독립이라고 가정할 수 있다. $\eta_i = logit(\pi_i)$라 할때, i번째 선형 예측값 $\hat{\eta}_i$의 CMSEP는

$$E[(\hat{\eta}_i - \eta_i)^2 | y_i, \alpha, \sigma^2] = Var(\eta_i|y_i, \alpha, \sigma^2) + C_i(y_i, \alpha, \sigma^2)$$

로 표현되며, 이때 $C_i(y_i, \alpha, \sigma^2)$는 비음(nonnegative)의 값을 가지는 편향항(bias term)으로 $O_p(n^{-1})$에 따른다. 이 경우, α와 σ^2항들이 알려져 있으면, 위 CMSEP는 $Var(\eta_i|y_i, \alpha, \sigma^2)$에 수렴한다. 이와는 별도로, 일반적으로 쓰이는 η_i에 대한 "예측의 무조건부 평균 제곱 오차(Unconditional Mean Square Error of Prediction: UMSEP)"는, 관찰값 y_i에 의존하지 않는,

$$E[(\hat{\eta}_i - \eta_i)^2 | \alpha, \sigma^2]$$

이 되며, 따라서 UMSEP는, $Var(\eta_i|y_i, \alpha, \sigma^2)$이 관찰값 y_i에 의존하지 않는 경우를 제외하고는, $Var(\eta_i|y_i, \alpha, \sigma^2)$에 확률 수렴하지 않는다. 선형 혼합 모형에서 CMSEP와 UMSEP가 일치하게 되는 이유는 바로 위의 $Var(\eta_i|y_i, \alpha, \sigma^2)$가 y_i에 의존하지 않기 때문이다. Booth와 Hobert(1998)는 $Var(\eta_i|y_i, \alpha, \sigma^2)$의 추정값을 얻기 위해 Laplace 근사를 이용하였고 – 이에 대해서는 상기한 다른 수치 적분 방법이나 Monte-Carlo 기법을 이용할 수도 있다 – $C_i(y_i, \alpha, \sigma^2)$항은 $Var(\eta_i|y_i, \hat{\alpha}, \hat{\sigma}^2)$에 대한 1차 Taylor 확장을 사용하여 $\boldsymbol{A}_i \boldsymbol{I}^{-1}(\alpha, \sigma^2) \boldsymbol{A}_i$로 근사하였다. 여기서 행렬 $\boldsymbol{A}_i$는 Taylor 확장으로부터 얻어질 수 있고, $\boldsymbol{I}(\alpha, \sigma^2)$은 정보 행렬이다. 이때 모수 α와 σ^2은 알려져 있지 않기 때문에 이들의 추정값을 이용하여 CMSEP의 추정값을 계산할 수 있다. 즉

$$\begin{aligned} CMSEP &\approx [Var(\eta_i|y_i, \alpha, \sigma^2) + \boldsymbol{A}_i \boldsymbol{I}^{-1}(\alpha, \sigma^2) \boldsymbol{A}_i]|_{\hat{\alpha}, \hat{\sigma}^2} \\ &= \hat{V}_i + \text{보정항} \end{aligned}$$

이고, 여기서 $\hat{V}_i$은 바로 위의 순수(naive) 추정값 $Var(\eta_i|y_i, \hat{\alpha}, \hat{\sigma}^2)$이며, 보정항(correction term)은 $C_i(y_i, \alpha, \sigma^2)$의 추정값 $\boldsymbol{A}_i(\hat{\alpha}, \hat{\sigma}^2) \boldsymbol{I}^{-1}(\hat{\alpha}, \hat{\sigma}^2) \boldsymbol{A}_i(\hat{\alpha}, \hat{\sigma}^2)$이다. 즉 다시 말하면 CMSEP의 두번째 항 $C_i(y_i, \alpha, \sigma^2)$은 모수 α와 σ^2의 추정값을 사용함으로 인해서 순수 추정값이 설명

하지 못하는 실제 분산의 일부분을 보정해 주는 역할을 한다고 볼 수 있는 것이다. 한편 Booth와 Hobert(1998)는 이 보정항의 차수(order)가 $Var(\eta_i|y_i, \alpha, \sigma^2)$을 추정할 때 생기는 편향의 차수와 같다고 보고, 따라서 추정의 정확도가 요구되는 경우 이 편향은 무시될 수가 없기 때문에, CMSEP의 추정값으로

$$CMSEP \approx \hat{V}_i + \text{보정항} + \text{편향} \tag{4.2}$$

을 제시하였다. 여기서 편향, $b_i(\hat{\alpha}, \hat{\sigma}^2; y_i)$,은 $\hat{V}_i$에 대한 2차 Taylor 확장의 2차항 값에대한 추정값이며, 이를 얻기 위해 Booth와 Hobert(1998)는 자료가 주어진 조건하에서 얻어지는 조건부 편향에 대한 모수적 붓스트랩 추정값을 사용하였다. 결론적으로 예측값의 표준 오차에 대한 정확도가 요구되는 경우에는 그 추정값을 얻기 위해 붓스트랩 추정 방법을 – 비록 구하기가 쉽지 않더라도 – 사용할 필요가 있으며, 그렇지 않은 경우에는 위의 순수 추정값만으로도 충분하다고 보았다.

5. 현재의 연구 진행 방향 및 GLMM의 장단점

GLMM에 대한 현재의 연구 진행 방향으로는 MCMC 또는 EM 알고리즘을 혼합한 Monte-Carlo 방식등을 통하여 모형을 적합시키는 문제, Gibbs sampler를 이용하여 모수를 추정하는 문제(McCulloch, 1994, 1997; Hobert와 Casella, 1996)등이 연구되고 있으며, 명목 자료나 순서 자료를 반응값으로 가지는 경우에 대한 모형 적합방법의 확장등도 연구되고 있다 (Hedeker와 Gibbons, 1994). 한편, Bayesian 측면에서의 모형 적합에 관한 연구도 실시되었다 (Gilks, et al., 1993; Zeger와 Karim, 1991). 사실 Bayesian의 경우, 추정에 있어서 모수 효과와 변량 효과에 대한 구분을 따로 하지 않기 때문에 "혼합" 효과 모형을 설정하는 것이 상황을 더 복잡하게 만드는 것은 아니지만, Natarajan과 McCulloch(1995)는 분산 성분에 대해 flat prior를 사용하는 것은 바람직하지 않다고 지적하였다.

현재 GLMM이 가지는 문제점 중의 하나는 그 모형이 변량효과에 대한 정규성 가정에 기초하고 있다는 점이다. 따라서 비모수 변량 효과를 가정하는 경우의 연구도 진행되고 있으며(Agresti, 1993; Aitkin과 Francis, 1995), 이와 더불어 GEE를 사용하는 연구도 진행되고 있다 (Fitzmaurice, 1995; Lipsitz, et al., 1994). 또한 GLMM은, 상대적으로 넓은 지역에서 상대적으로 적은 수의 관찰값이 얻어질 때 지역-중심(area-specific) 관심 모비율을 추정하기 위해서, 변량 효과 모형을 사용함으로써 추정값의 효율을 증대시키는 소-지역 추정법(small-area estimation)에서도 효과적으로 응용된다 (Booth, 1995; Ghosh와 Rao, 1994; Ghosh, et al., 1998). 공변량에 대한 측정 오차가 있는 경우의 편향 및 분산 성분 검정 문제 또한 최근의 관심 사항이다 (Lin과 Carroll, 1999; Wang, et al., 1998).

GLMM은 다른 여러가지 모형들, 예를 들면 계층구조 일반화된 선형 모형(HGLM), 베타-이항 모형, 포아송-감마 모형, 완전 베이지안 모형(Full Bayesian Model), 그리고 조건부 최대 가능도 모형(Conditional Maximum Likelihood Model)등에 비해 보다 효과적으로 자

료를 분석할 수 있는 방법이다. 즉 GLMM을 통하여 변량 효과에 대한 정규성을 가정함으로서, 변량 효과내의 복잡한 상호 의존성을 효과적으로 모형화할 수 있게 되었으며, 이는 다른 모형적 방법들이 해결할 수 없었던 점이다. 또한 여러가지 연결 함수를 사용하여 다양한 자료를 분석할 수 있다는 점도 GLMM의 장점중 하나라 하겠다. 하지만 가장 큰 단점으로는 우선 계산상의 복잡함을 들 수 있다. GLMM을 분석하기 위한 알고리즘들 중 어느 것이 최선의 방법인지는 아직 의견의 통일을 보이지 않고 있다. 더군다나 이 모든 알고리즘은 계산상 요구되는 컴퓨터의 용량이 워낙 크기 때문에 특히 대형 자료나 여러 개의 변량 효과를 포함하고 있는 자료의 분석에 어려움이 있다.

6. 앞으로의 연구 가능성

다음은 GLMM과 그에 관련된 주제를 가지고 열린 미국 NSF-CBMS 지역회의에서 논의되었던 앞으로의 연구 가능성에 관한 주제들이다 (McCulloch, 1999).

1. 모형 적합 방법 개발 문제:
   변량 효과가 첨가된 모형에서 ML 추정값을 얻기란 쉽지 않고, 이를 위해 개발된 알고리즘들 중 일부는 근사가 잘 이루어지지 않는다. 고차원 적분을 필요로 하는 복잡한 모형의 경우 수치 적분을 통한 가능도 함수를 계산하기란 일반적으로 쉽지 않다. 이러한 경우에 적절한 모형 적합 방법 개발이 요구된다. BUGS등을 이용한 Bayesian 측면에서의 분석도 진행되고 있지만 사전확률 선택의 효과는 아직도 연구 대상이다.

2. 연결 함수의 선택과 진단 문제:
   어떤 연결 함수를 선택할 것이며, 그 선택이 제대로 된 것인지에 대한 진단은 어떻게 할 수 있을 것인가? 즉 연결 함수들의 혼합적 사용 가능성, 연결 함수족들에 대한 연구, 비모수적 연결 함수의 개발, 그래프적 진단 방법등에 관한 연구가 필요하다. 이와 관련된 기존의 연구들로 Li와 Duan(1989), Mallick과 Gelfand(1994), Pregibon(1980), Weisberg와 Welsh(1994), 그리고 Xie, Simpson과 Carroll(1997)등을 참조하기 바란다.

3. 변량 효과의 분포에 대한 선택 및 진단의 문제:
   어떤 분포를 사용할 것인가? 그 선택이 적절한 것인지는 어떻게 판단할 수 있는가? 변량 효과란 눈으로 확인될 수 없는데 분포의 선택 여부가 과연 중요한가? (그렇다는 측면의 McCulloch(1997)의 연구가 있는 반면, 그렇지 않다는 측면의 Neuhaus, Hauck와 Kalbfleisch(1992)의 연구가 있다.) 변량 효과 $\boldsymbol{u}_i$의 추정값을 이용한 그래프가 분포 선택에 효과적인가? (그렇다는 측면의 Lange과 Ryan(1989)의 연구가 있는 반면, 그렇지 않다는 Verbeke과 Lesaffre(1996)의 연구도 있다.)

4. 이상치의 판단과 잔차항의 분석 문제:
   이상치의 구별은 어떻게 해야 하며, 그것이 가지는 효과는 어떻게 판단할 것인가? 선

형 혼합 모형(LMM)의 경우에도 이상치에 관한 연구는 Christensen, Pearson과 Johnson(1992), Hodges(1998)등 소수의 연구만 이루어져 있다. 사실 혼합 모형의 경우에는 자료 제거(data deletion)가 경우별로 구분되어야 하며(예를 들면, 각 개체내 반복 자료중 하나를 제거하는 것과 한 개체의 자료 자체를 제거하는 것), 이 경우 이상치의 여부를 조사하기 위해서 각 자료는 서로 다른 분산을 가지고 측정되어야 한다. GLMM의 경우에는 아직 이에 대한 연구가 많이 이루어지고 있지 않은 상황이다.

5. 소표본에서의 추론 및 분산 성분의 검정 문제:
   추정이나 검정, 신뢰구간에 관한 GLMM의 모든 연구는 대표본에 근거한 방법론에 의해 진행되어 왔다. 그러면 과연 표본수가 충분히 크다는 것은 어느정도를 의미하는가? 소표본의 경우 모수 효과와 변량 효과에 대한 검정의 문제는, 특히 변량 효과의 경우, 관심 사항중 하나가 아닐 수 없다. 분산 성분 검정에 관한 연구로는 Lin(1997), Jiang(1998)등을 참조하기 바란다.

6. 예측값의 정확도 문제:
   예측된 값들이 과연 얼마나 정확한가? 서로 다른 예측값들간의 검정은 어떻게 이루어질 수 있는가? 변량 효과의 최량 예측값(best predicted value)을 얻기 위해 모수 효과들의 추정값을 사용하면, 그 효과는 어떻게 되는가?(예: Peixoto와 Harville, 1986; Natarajan과 McCulloch, 1999) 예측 오차(prediction error)를 어떻게 계산할 것인가?(예: Booth와 Hobert, 1998)등에 관한 연구가 필요하다.

7. 결론

일반화된 선형 혼합 모형(GLMM)은 선형 혼합 모형(LMM)과 일반화된 선형 모형(GLM)의 틀들을 연결하여, 정규 연속형 자료뿐만 아니라, 비정규 자료, 비선형 모형 자료, 변량 효과를 지닌 공분산 구조 자료, 계수 형태로 나타나는 범주형 자료들에 대해서도 단순하고 융통성 있게 분석을 할 수 있는 통합된(unified) 모형 적합 방법이다. 본 연구에서는 이에 대한 개요와 더불어, 모형 적합을 위해 제시된 기법들중 의사가능도(QL)를 이용한 추정 방법과 Monte-Carlo를 이용한 추정 방법들에 대해 알아보았다. 이 외에도, 특수한 상황하에서 이용될 수 있는 베타-이항 모형이나 포아송-감마 모형, 또 GEE를 사용하는 방법, 베이즈 방법을 이용한 적용등이 가능하나, 이에 대해서는 본 연구에서 다루지 않았다. 계산상의 복잡성이 해결된다면, GLMM은 기존의 통계적 방법론들에 비해, 보다 효율적인 도구가 될 것이며, 모형 진단의 문제, 모수들에 대한 검정 문제 및 신뢰 구간 설정 문제등이 아직 남아 있는 상태에서 GLMM의 연구 범위는 매우 넓다고 볼 수 있겠다.

참고문헌

[1] 이준영 (1999). 일반화된 선형 혼합 모형: 선형 혼합 모형과 일반화된 선형 모형의 연결, <응용 통계>, 고려대학교 통계연구소, 제 14권, 27-40.

[2] Agresti, A. (1993). Distribution-free fitting of logistic models with random effects of repeated categorical responses, *Statistics in Medicine*, vol. 12, 1969-1987.

[3] Aitkin, M. and Francis, B.J. (1995). Fitting overdispersed generalized linear models by nonparametric maximum likelihood, *GLIM Newsletter*, vol. 25, 37-45.

[4] Anderson, D.A. and Aitkin, M. (1985). Variance component models with binary responses: Interviewer variability, *Journal of the Royal Statistical Society, Series B.*, vol. 47, 203-210.

[5] Bock, R.D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm, (corr: vol. 47, 369), *Psychometrika*, vol. 46, 443-459.

[6] Booth, J.G. (1995). Bootstrap methods for generalized linear mixed models with applications to small area estimation, *Proceedings of the 10th International Workshop on Statistical Modelling*, vol. 104, 43-51.

[7] Booth, J.G. and Hobert, J.P. (1998). Standard errors of prediction in generalized linear mixed models, *Journal of the American Statistical Association*, vol. 93, 262-272.

[8] Booth, J.G. and Hobert, J.P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte-Carlo EM algorithm, *Journal of the Royal Statistical Society, Series B.*, vol. 61, 265-285.

[9] Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association*, vol. 88, 9-25.

[10] Breslow, N.E. and Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion, *Biometrika*, vol. 82, 81-91.

[11] Cnaan, A., Laird, N.M. and Slasor, P. (1997). Using a generalized linear mixed model to analyse unbalanced repeated measures and longitudinal data, *Statistics in Medicine*, vol. 16, 2349-2380.

[12] Chan, J.S.K. and Kuk, A.Y.C. (1997). Maximum likelihood estimation for probit-linear mixed models with correlated random effects, *Biometrics*, vol. 53, 86-97.

[13] Christensen, R., Pearson, L.M. and Johnson, W. (1992). Case-deletion diagnostics for mixed models, *Technometrics*, vol. 34, 38-45.

[14] Daniels, M.J. and Gatsonis, C. (1999). Hierarchical generalized linear models in the analysis of variations in health care utilization, *Journal of the American Statistical Association*, vol. 94, 29-42.

[15] Fitzmaurice, G.M. (1995). A caveat concerning independence estimating equations with multivariate binary data, *Biometrics*, vol. 51, 309-317.

[16] Ghosh, M. and Rao, J.N.K. (1994). Small area estimation: An appraisal (with discussion), *Statistical Science*, vol. 9, 55-93.

[17] Ghosh, M., Natarajan, K., Stroud, T. and Carlin, B. (1998). Generalized linear models for small area estimation, *Journal of the American Statistical Association*, vol. 93, 273-282.

[18] Gibbons, R.D., Hedeker, D., Charle, S.C. and Frisch, P. (1994). A random-effects probit model for predicting medical malpractice claims, *Journal of the American Statistical Association*, vol. 89, 760-767.

[19] Gilks, W.R., Wang, C.C., Yvonnet, B. and Coursaget, P. (1993). Random-effects models for longitudinal data using Gibbs sampling, *Biometrics*, vol. 49, 441-453.

[20] Gilmour, A.R., Anderson, R.D. and Rae, A.L. (1985). The analysis of binomial data by a generalized linear mixed model, *Biometrika*, vol. 72, 593-599.

[21] Goldstein, H. (1991). Nonlinear multilevel models, with an application to discrete response data, *Biometrika*, vol. 78, 45-51.

[22] Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems, *Journal of the American Statistical Association*, vol. 72, 320-340.

[23] Harville, D.A. and Mee, R.W. (1984). A mixed-model procedure for analyzing ordered categorical data, *Biometrics*, vol. 40, 393-408.

[24] Hedeker, D. and Gibbons, R.D. (1994). A random-effects ordinal regression model for multilevel analysis, *Biometrics*, vol. 50, 933-944.

[25] Hobert, J.P. and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierachical linear mixed models, *Journal of the American Statistical Association*, vol. 91, 1461-1473.

[26] Hodges, J.S. (1998). Some algebra and geometry for hierachical models, applied to diagnostics (with discussion), *Journal of the Royal Statistical Society, Series, B.*, vol. 60, 497-536.

[27] Jiang, J. (1998). Consistent estimators in generalized linear mixed models, *Journal of the American Statistical Association*, vol. 93, 720-729.

[28] Kachman, S.D. and Stroup, W.W. (1994). Generalized linear mixed models: an application, *Proceedings of the Kansas State University of Applied Statistics in Agriculture*, 99-111.

[29] Laird, N.M. (1978). Empirical Bayes methods for two-way contingency tables, *Biometrika*, vol. 65, 581-590.

[30] Langford, I.H. (1994). Using a generalized linear mixed models to analyze dichotomous choice contingent variation data, *Land Economics*, vol. 70, 507-514.

[31] Lange, N. and Ryan, L. (1989). Assessing normality in random effects models, *Annals of Statistics*, vol. 17, 624-642.

[32] Lee, Y. and Nelder, J.A. (1996). Hierachical generalized linear models (with discussion), *Journal of the Royal Statistical Society, Series, B.*, vol. 58, 619-679.

[33] Li, K.-C. and Duan, N. (1989). Regression analysis under link violation, *Annals of Statistics*, vol. 17, 1009-1052.

[34] Lin, X. (1997). Variance component testing in generalized linear mixed models with random effects, *Biometrika*, vol. 84, 309-326.

[35] Lin, X. and Carroll, R.J. (1999). SIMEX variance component tests in generalized linear mixed measurement error models, *Biometrics*, vol. 55, 613-619.

[36] Lipsitz, S.R., Fitzmaurice, G.M., Orav, E.J. and Laird, N.M. (1994). Performance of generalized estimating equations in practical situations, *Biometrics*, vol. 50, 270-278.

[37] Littell, R.C., Milliken, G.A., Stroup, W.W. and Wolfinger, R.D. (1996). *SAS System for Mixed Models*, SAS Institute Inc., Cary, North Carolina.

[38] Liu, Q. and Pierce, D.A. (1994). A note on Gauss-Hermite quadrature, *Biometrika*, vol. 81, 624-629.

[39] Mallick, B.K. and Gelfand, A.E. (1994). Generalized linear models with unknown link functions, *Biometrika*, vol. 81, 237-245.

[40] McCullagh, P. and Nelder, J. (1989). *Generalized linear models, 2nd edition*, Chapman & Hall, London.

[41] McCulloch, C.E. (1994). Maximum likelihood variance components estimation for binary data, *Journal of the American Statistical Association*, vol. 89, 330-335.

[42] McCulloch, C.E. (1997). Maximum likelihood algorithms for generalized linear mixed models, *Journal of the American Statistical Association*, vol. 92, 162-170.

[43] McCulloch, C.E. (1999). *NSF/CBMS Regional conference on generalized linear mixed models and related topics*, June 8-12, Department of Statistics, University of Florida, Gainesville, FL, USA.

[44] Montgomery, M.R., Richards, T. and Braun, H.I. (1986). Child health, breast-feeding, and survival in Malaysia: A random-effects logit approach, *Journal of the American Statistical Association*, vol. 81, 297-309.

[45] Murphy, D.M. and Wang, D. (1998). Family and sociodemographic influences on patterns of leaving home in postwar Britain, *Demography*, vol. 35, 293-305.

[46] Natarajan, R. and McCulloch, C.E. (1995). A note on existence of the posterior distribution for a class of mixed models for binomial responses, *Biometrika*, vol. 82, 639-643.

[47] Natarajan, R. and McCulloch, C.E. (1999). Modeling heterogeneity in nested survival data, *Biometrics*, vol. 55, 553-559.

[48] Nee, V. (1996). The emergence of a market society: Changing mechanisms of stratification in China, *American Journal of Sociology*, vol. 101, 908-949.

[49] Neuhaus, J.M., Hauck, W.W. and Kalbfleisch, J.D. (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models, *Biometrika*, vol. 79, 755-762.

[50] Neuhaus, J.M., Kalbfleisch, J.D. and Hauck, W.W. (1994). Conditions for consistent estimation in mixed-effects models for binary matched-pairs data, *The Canadian Journal of Statistics*, vol. 22, 139-148.

[51] Patterson, H.D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal, *Biometrika*, vol. 58, 545-554.

[52] Peixoto, J. and Harville, D.A. (1986). Comparisons of alternative predictors under the balanced one-way random model, *Journal of the American Statistical Association*, vol. 81, 431-436.

[53] Pinheiro, J.C. and Bates, D.M. (1995). Approximations to the log-likelihood function in the non-linear mixed-effects model, *Journal of Computational and Graphical Statistics*, vol. 4, 12-35.

[54] Prosser, R., Rasbash, J. and Goldstein, H. (1995). *MLn software for multilevel analysis*, Institute of Education, University of London, London, UK.

[55] Pregibon, D. (1980). Goodness of link tests for generalized linear models, *Applied Statistics*, vol. 29, 15-24.

[56] Rasch, G. (1961). On general laws and the meaning of measurement in psychology, *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 4, 321-333.

[57] Robinson, G.K. (1991). That BLUP is a good thing: The estimation of random effects (with discussion), *Statistical Sciences*, vol. 6, 15-51.

[58] Sampson, R.J., Raudenbush, S.W. and Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy, *Science*, vol. 227, 918-924.

[59] SAS Institute Inc. (1999). *Chapter 3. The NLMIXED procedure (Draft)*, Cary, NC: SAS Institute Inc.

[60] Searle, S.R., Casella, G. and McCulloch, C.E. (1992). *Variance Components*, John Wiley & Sons, New York.

[61] Stiratelli, R., Laird, N. and Ware, J.H. (1984). Random-effects models for serial observations with binary responses, *Biometrics*, vol. 40, 961-971.

[62] Stroup, W.W. and Kachman, S.D. (1994). Generalized linear mixed models - An overview, *Proceedings of the Kansas State University Conference of Applied Statistics in Agriculture*, 82-98.

[63] Wang, N., Lin, X., Gutierrez, R.G. and Carroll, R.J. (1998). Bias analysis and SIMEX approach in generalized linear mixed measurement error models, *Journal of the American Statistical Association*, vol. 93, 249-261.

[64] Wei, G.C.G. and Tanner, M.A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms, *Journal of the American Statistical Association*, vol. 85, 699-704.

[65] Weisberg, S. and Welsh, A.H. (1994). Adapting for the missing link, *Annals of statistics*, vol. 22, 1674-1700.

[66] Wolfinger, R.D. and O'Connell, M. (1993). Generalized linear mixed models: A pseudo-likelihood approach, *Journal of Statistical Computation and Simulation*, vol. 48, 233-243.

[67] Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population, *Journal of the American Statistical Association*, vol. 91, 217-221.

[68] Xie, M., Simpson, D.G. and Carroll, R.J. (1997). Scaled link functions for heterogeneous ordinal response data. In: *Modelling longitudinal and spatially correlated data: Methods, applications and future directions*, T.G. Gregoire (Ed.), Springer-Verlag, New York.

[69] Zeger, S.L. and Karim, M.R. (1991). Generalized linear models with random effects: a Gibbs sampling approach, *Journal of the American Statistical Association*, vol. 86, 79-86.

[70] Zeger, S.L., Liang, K.-Y. and Albert, P.S. (1988). Models for longitudinal data: A generalized estimating equation approach (corr. vol. 45, 347), *Biometrics*, vol. 44, 1049-1060.

[71] Zou, X.-H., Perkins, A.J. and Hui, S.L. (1999). Comparisons of software packages for generalized linear multilevel models, *The American Statistician*, vol. 53, 282-290.

[1999년 11월 접수, 2000년 8월 채택]

A Study for Recent Development of Generalized Linear Mixed Model

Juneyoung Lee[1)]

ABSTRACT

The generalized linear mixed model framework is for handling count-type categorical data as well as for clustered or overdispersed non-Gaussian data, or for non-linear model data. In this study, we review its general formulation and estimation methods, based on quasi-likelihood and Monte-Carlo techniques. The current research areas and topics for further development are also mentioned.

Keywords: Linear model; Mixed model; Quasi-likelihood; Monte-Carlo simulation.

1) Research Fellow, Department of Preventive Medicine and Institute for Environmental Health, College of Medicine, Korea University. E-mail: jyleeuf@mail.korea.ac.kr