The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.2.123
28(2), 123–136

Review of Mixed-Effect Models

Youngjo Lee[a,1]

[a]Department of Statistics, Seoul National University

(Received April 16, 2015; Revised April 23, 2015; Accepted April 23, 2015)

Abstract

Science has developed with great achievements after Galileo's discovery of the law depicting a relationship between observable variables. However, many natural phenomena have been better explained by models including unobservable random effects. A mixed effect model was the first statistical model that included unobservable random effects. The importance of the mixed effect models is growing along with the advancement of computational technologies to infer complicated phenomena; subsequently mixed effect models have extended to various statistical models such as hierarchical generalized linear models. Hierarchical likelihood has been suggested to estimate unobservable random effects. Our special issue about mixed effect models shows how they can be used in statistical problems as well as discusses important needs for future developments. Frequentist and Bayesian approaches are also investigated.

Keywords: Mixed effect models, hierarchical generalized linear models, hierarchical likelihood, random effects.

1. 서론

통계학은 기본적으로 관측이 가능한 반응변수 y와 설명변수 x의 관계를 미지의 모수 θ를 통해서 표현한다. Bayes (1763)에서는 θ에 대한 사전 확률 분포를 가정하고 확률(probability)을 사용하여 θ에 대한 추론이 가능함을 보였다. 그리고 R. A. Fisher는 1921년 확률과는 다른 개념인 가능도(likelihood) 함수를 제시하였으며, 이를 통해서 θ에 대한 사전 확률 분포에 대한 가정 없이도 모수에 대한 통계적 추론이 가능함을 보였다. Birnbaum (1962)를 보면 자료에 있는 모수 θ에 대한 모든 정보는 가능도에 있다는 가능도 원칙(likelihood principle)를 증명하여, 가능도를 통계추론에 사용하는 이론적 기반을 마련하였다.

1861년 영국의 천문학자 Airy은 항성간의 거리나 행성의 지름 등을 측정하는 천문학 문제를 해결하기 위하여 기존의 일원배치 모형에 특정 밤의 대기나 조건을 추가하였는데, 이는 '변량 효과(random effect)'라는 새로운 개념을 도입한 최초의 혼합효과모형을 사용한 것으로 알려져 있다. 이러한 혼합모형에서는 미지의 모수 θ뿐 아니라, 관측할 수 없는 변량효과 v에 대한 추론이 요구되는데, 관측 가능

This research was supported by an NRF grant funded by Korea government (MEST) (No. 2011-0030810) and the Brain Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (2014M3C7A1062896).
[1]Department of Statistics, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul 151-742, Korea.
E-mail: youngjo@snu.ac.kr

한 변량들로만 이루어진 Fisher의 가능도 만으로는 관측할 수 없는 변량에 대한 추론이 불가능하다. Pearson (1920)은 Fisher의 가능도 방법으로는 관측되지 않은 미래의 변량을 예측할 수 없는 한계를 지적하였고, 베이지안 추론법이 관측할 수 없는 변량효과를 분석하기 위한 유일한 방법이라고 주장하였다. 그러나, 미지의 모수 θ에 대한 사전 확률 분포의 가정 없이 관측할 수 없는 변량효과 v에 대한 추론을 하기 위하여, Lauritzen (1974), Berger와 Wolpert (1984), Butler (1986), Bayarri 등 (1988), Bjørnstad (1996) 등과 같은 학자들이 여러 가지 방법들을 제시해왔다. 관측되지 않는 변량 v를 추론하기 위해서는 관측 가능한 변량만을 포함하는 기존의 가능도를 확장해야한다. Lee와 Nelder (1996)는 혼합선형모형을 확장한 다단계 일반화 선형모형(hierarchical generalized linear models; HGLMs)을 제안한 후, 다단계 가능도(hierarchical likelihood; h-likelihood)를 제시하여 모수에 대한 사전분포의 가정 없이도 관측되지 않는 변량의 추론이 가능함을 보였다. Bjørnstad (1996)는 자료에 있는 미지의 모수 θ와 관측할 수 없는 변량효과 v에 대한 모든 정보가 다단계 가능도에 있다는 확장된 가능도 원칙(extended likelihood principle)을 증명하여, 다단계 가능도를 통계적 추론에 사용하는 이론적 발판을 마련하였다. Pawitan (2001)은 다단계 가능도가 관측 못하는 변량을 추론하기 위한 기존 가능도의 타당한 확장이라고 역설하였다. 따라서, 다단계 가능도를 통해 미지의 모수 θ에 대한 추론뿐만 아니라 관측되지 않는 변량 v까지 모두 추론할 수 있게 되었다.

최근 혼합효과모형은 일반화 선형혼합모형, 다단계 일반화선형모형 등 여러 다양한 모형으로 확장되고 새로운 추론법들이 개발되어 이에 대한 관심이 고조되고 있다. 이러한 추세에 부응하여, 한국통계학회의 응용통계연구 28권 2호는 그 중요성을 인식하여 혼합효과모형의 특집호를 준비하였다. 특집호에 실린 총 22편 논문들은 기존 혼합효과모형, 베이지안 방법론, 다단계 가능도 방법론 등에 대한 리뷰 논문들을 포함하여, 경시적 자료, 시공간 자료, 이미지, 약동학, 기후, 품질관리, 금융, 유전학, 의학, 농학 등 다양한 분야에 대한 이론과 응용, 계산방법에 대한 연구, 그리고 한국에서 자체적으로 개발한 변량효과모형들에 대한 통계 패키지 소개 등이 있다.

자료를 측정하고, 저장하고, 처리하는 기술이 발달하면서 우리가 접할 수 있는 자료는 매우 커지고, 새로운 형태들이어서 요구되는 통계 모형 또한 매우 복잡, 다양해지고 있다. 혼합효과모형은 이러한 빅데이터 시대에 복잡한 자료들을 잘 설명하는 모형이다. 특히 다단계 가능도는 기존 베이지안 확률과 Fisher의 가능도를 포괄하는 새로운 통계 패러다임으로 새로이 확장된 모형들에 사용할 수 있는 획기적 방법으로서, 한국에서 이 분야에 대한 연구를 국제적으로 선도하고 있다. 혼합효과모형 특집호 논문들이 한국통계학회 회원들에게 혼합효과모형에 대한 이해와 안목을 넓히고, 아울러 다양한 분야들에서 혼합효과모형이 사용되는 계기가 되기를 바란다.

이 논문에서는 관측되지 않는 변량 v에 대한 추론의 중요성을 간단한 예시를 통해 (2절), 혼합선형모형을 확장한 다단계 일반화 선형 모형과 더 확장된 모형들과 방법론들에 대해 살펴보고 (3절), 이와 관련된 여러 중요한 통계 문제들을 생각해보고자 한다 (4절). 확장된 가능도인 다단계 가능도는 하나의 새로운 통계 패러다임이므로 이에 대한 리뷰는 통계 전 분야에 걸쳐서 해야 하나, 지면 관계상 현재 개발되고 있는 중요한 분야들에 대하여 우선적으로 다룬다. 기존의 베이지안 방법론은 잘 알려져 있으므로, 이에 대해서는 베이지안 리뷰 논문인 Lee 등 (2014)을 참고하기 바란다.

2. 변량 효과의 필요성

2.1. 미래 관측치의 예측 문제

어떤 한 환자의 지난 5주간 간질성 발작의 횟수를 $y = (3, 2, 5, 0, 4)$라고 하자. 이 환자의 발작 횟수를 서로 독립이고 평균 모수 θ를 가지는 포아송 분포를 따른다고 하고, 이 환자의 다음주 발작 횟수 v의 예

측 확률 분포를 구해보자. Fisher의 가능도는 다음과 같이 나타낼 수 있으며,

$$f_\theta(3,2,5,0,4) = \frac{\exp(-5\theta)\theta^{3+2+5+0+4}}{3!\,2!\,5!\,0!\,4!}.$$

이를 이용한 θ의 최대가능도 추정값

$$\hat{\theta} = \frac{3+2+5+0+4}{5} = 2.8$$

을 구하고, 포아송 분포의 평균 모수 θ에 $\hat{\theta}$ 대체하는 삽입법(plug-in technique)을 이용하여 v의 예측 분포를

$$f_{\hat{\theta}}(v=i|y) = f_{\hat{\theta}}(v=i) = \frac{\exp(-2.8)2.8^i}{i!}, \quad i=0,1,\ldots$$

로 구할 수 있다. 하지만, Pearson (1920)은 이렇게 구하는 Fisher의 가능도에 의한 v 예측의 한계로, θ를 추정함에 있어서 존재하는 불확실성을 v의 확률분포에 반영하지 못하는 것을 지적하였다. 이 때문에 Pearson은 v의 예측확률을 구할 때는 θ의 사전 확률 분포 $\pi(\theta)$를 가정하고 θ를 적분을 통해 소거하므로서 θ의 추정에 따른 불확실성을 고려하는 베이지안 방법을 사용해야 함을 주장하였다. Pearson은 Jeffrey의 사전확률 $\pi(\theta) \propto \theta^{-1/2}$를 이용하여 주어진 자료에 대한 v의 예측 확률을 사후확률로 구하였다:

$$P(v=i|y) \propto \frac{(i+3+2+5+0+4+0.5)!}{i!\,6^{i+3+2+5+0+4+0.5}}.$$

한편, 이 문제에서 다단계 가능도는

$$f_\theta(3,2,5,0,4,v) = \frac{\exp(-6\theta)\theta^{3+2+5+0+4+v}}{3!\,2!\,5!\,0!\,4!\,v!}$$

와 같이 나타낼 수 있다. 이 다단계 가능도를 이용하여 v의 함수로써 최대 다단계 가능도 추정치

$$\hat{\theta}(v) = \frac{3+2+5+0+4+v}{6}$$

를 구한 뒤, θ를 $\hat{\theta}(v)$로 대체한 프로파일 가능도

$$f_{\hat{\theta}(v)}(3,2,5,0,4,v)$$

를 통해 v에 대한 예측 확률 분포 함수를 구할 수 있다. Pearson이 주장한 베이지안 예측법은 θ에 대한 사전 확률 분포를 가정해야하나, 예시한 바와 같이 다단계 가능도를 이용한 추정방법은 사전 확률 분포의 가정 없이도 θ의 추정에 따른 불확실성을 고려한 v에 대한 예측 확률 분포를 구할 수 있음을 알 수 있다 (Figure 2.1).

2.2. 다중 N 가설검정 문제에서, 관측할 수 없는 변량들

다중검정(multiple testing) 문제에서 Benjmini와 Hochberg (1995)는 오발견율(false discovery rate; FDR)이라는 측도를 제안하였다. 예를 들어, 총 N개의 검정 대상 중에, 어떤 통계적 방법론을 통해 R개의 가설들을 기각했다고 하면, 다음 Table 2.1을 구할 수 있다. 여기서 Benjmini와 Hochberg (1995)는 오발견율을 FDR $= E(v_{01}/R)$로 정의하였다. 이 때, Table 2.1에서 R과 N외에 $v_{00}, v_{01}, v_{10}, v_{11}$은 모두 관측할 수 없는 변량들임을 알 수 있다. 따라서, 다중 검정 문제에서도 올바르게 오발견율을 추정하고 문제를 풀기 위해서는 관측할 수 없는 변량이 포함된 통계모형과 추론이 필요함을 알 수 있다.

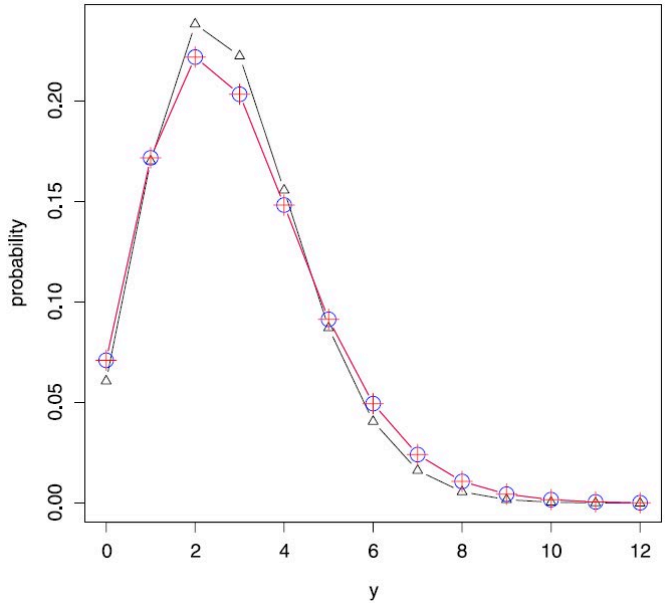

Figure 2.1. Predictive density of the number of seizure counts: Plug-in method (△), Bayesian method (◯) and h-likelihood method (+).

Table 2.1. Unobservables of interest in testing N hypotheses

	Declared as null	Declared as alternative	Total
Null	v_{00}	v_{01}	N_0
Alternative	v_{10}	v_{11}	N_1
Total	$N-R$	R	N

3. 혼합선형모형의 확장

3.1. 혼합선형모형

변량효과 v가 포함된 가장 간단한 모형으로 혼합 선형모형을 둘 수 있다. y는 N개의 반응변수로 이루어진 벡터를, X와 Z는 각각 미지의 상수인 고정효과 β와 관측하지 못한 변량효과 v의 모형 행렬로서 $N\times p$, $N\times q$의 차원을 가진다고 하자. 이 때, 기본적인 혼합 선형모형은 다음과 같이 나타낼 수 있다.

$$y = X\beta + Zv + e. \tag{3.1}$$

이 때, 변량 v와 오차 e의 분포 가정은 $e \backsim \mathrm{MVN}(0,\Sigma)$, $v \backsim \mathrm{MVN}(0,D)$이며 보통 v와 e는 독립이라고 가정한다.

v가 주어졌을때, 반응변수 y는 $E(y|v) = X\beta + Zv$를 평균으로, Σ를 분산으로 하는 정규분포를 따르며, v는 평균이 0이고 분산이 D인 정규분포를 따른다. 이 경우, 다단계 (로그-) 가능도는 다음과 같다.

$$\begin{aligned} h &= \log f(y,v) = \log f(y|v) + \log f(v) \\ &= -\frac{1}{2}\log|2\pi\Sigma| - \frac{1}{2}(y-X\beta-Zv)^t\Sigma^{-1}(y-X\beta-Zv) - \frac{1}{2}\log|2\pi D| - \frac{1}{2}v^tD^{-1}v. \end{aligned} \tag{3.2}$$

모수가 주어졌다고 할 때, h를 최대로 하는 v의 추정량은 다음과 같이 얻을 수 있다.

$$\hat{v} = \left(Z^t\Sigma^{-1}Z + D^{-1}\right)^{-1} Z^t\Sigma^{-1}(y - X\beta) = E(v|y). \tag{3.3}$$

3.2. 다단계 일반화 선형모형

기존의 일반화 선형모형 (Nelder와 Wedderburn, 1972)에 변량 효과를 추가한 일반화 선형 혼합모형(Generalized linear mixed models; GLMMs)은 관측 가능한 변량 y를 지수족으로 확장하고, 변량 효과를 정규 분포를 따르는 모형만 다룬다 (Breslow와 Clayton, 1993). 한편, Lee와 Nelder (1996)은 변량 효과의 분포 또한 지수족으로 확장한 다단계 일반화 선형모형(Hierarchical Generalized Linear Models; HGLMs)을 제안하고, 다음과 같이 정의하였다.

(i) 변량효과 u가 주어졌을 때, 반응변수 y의 모형으로 다음을 만족하는 일반화 선형모형을 가정한다.

$$E\left(y|u\right) = \mu \quad \text{and} \quad \text{var}\left(y|u\right) = \phi V\left(\mu\right),$$

여기서 ϕ는 산포모수를, $V(\ )$는 분산함수를 의미한다. 정준 모수 $\theta = \theta(\mu)$를 정의하면, 로그 가능도함수의 커널은 다음과 같이 주어진다.

$$\sum \frac{y\theta - b(\theta)}{\phi}.$$

평균 μ는 연결함수 $g(\ )$를 통해, 다음과 같이 모형화 되며

$$\eta = g\left(\mu\right) = X\beta + Zv. \tag{3.4}$$

이때, $v = v\left(u\right)$는 변량효과 u의 단조 변환된 변량효과, β는 고정효과를 의미한다.

(ii) 변량효과 u는 모수 λ를 가지는 지수족 분포를 따른다.

예를 들어, $y|u$가 다음과 같은 평균을 가지는 포아송 분포를 따른다고 가정해보자.

$$\mu = E(y|u) = \exp(X\beta)u$$

로그 연결함수는 다음과 같은 선형식을 만든다.

$$\eta = \log\mu = X\beta + v$$

이때, $v = \log u$이고, u가 감마 분포를 따르면, v는 로그-감마 분포를 따르게 되며, 이러한 모형을 포아송-감마 다단계 일반화 선형모형이라고 한다. 일반화 선형 혼합모형의 경우, v의 분포로 정규 분포만 고려하게 된다. 다단계 일반화 선형모형의 특수한 예들로 $y|u$와 u의 분포별로 주로 많이 쓰이는 모형들은 Table 3.1와 같이 나타낼 수 있다. 모형을 적합 하는데 필요한 모수 및 변량효과들을 추정하기 위해, Lee와 Nelder (1996)은 다음과 같은 다단계 가능도(h-likelihood)를 정의한다.

$$h = \log f_\phi(y|v) + \log f_\lambda(v), \tag{3.5}$$

여기서 $\log f_\phi(y|v)$는 산포모수 ϕ를 가지는 y의 로그 조건부 확률 밀도 함수를 의미하고, $\log f_\lambda(v)$는 또 다른 산포모수 λ를 가지는 v의 로그 확률밀도 함수를 의미한다. 위의 모형 (3.4)에서 추정해야할 대상은 변량효과 v, 평균 모수 β, 산포 모수 $\tau = (\phi, \lambda)$이다.

Table 3.1. Examples of HGLMs with distributions of $y|v$ and u

$y\|v$ distribution	$g(\mu)$	u distribution	$v(u)$	Model
Normal	identity	Normal	identity	Linear mixed models
Binomial	logit	Beta	logit	beta-binomial model
Binomial	logit	Normal	identity	Binomial GLMM
Gamma	log	Normal	identity	Gamma GLMM
Gamma	log	Inverse-gamma	reciprocal	Gamma HGLM
Poisson	log	Normal	identity	Poisson GLMM
Poisson	log	Gamma	log	Poisson HGLM

또한 Lee와 Nelder (2001)은 다음과 같은 수정된 단면 가능도(adusted profile likelihood)를 정의하였다.

$$p_\alpha(l) = \left[l - \frac{1}{2}\log\left|\frac{D(l,\alpha)}{2\pi}\right| \right]\Bigg|_{\alpha=\tilde{\alpha}},$$

여기서 l은 로그 주변 가능도 또는 다단계 가능도를 의미하고, $D(l,\alpha) = -\partial^2 l/\partial\alpha\partial\alpha^T$이고 $\tilde{\alpha}$는 $\partial l/\partial\alpha = 0$의 해이다. 모형에 필요한 값들을 추정하는데 있어서 Lee와 Nelder (2001)은 변량효과 v는 식 (3.5)의 h를 최대화 시키는 값을, 평균 모수인 β는 수정된 단면 가능도 $p_v(h)$를, 분산 성분인 τ는 $p_{v,\beta}(h)$를 최대화 시키는 값으로 구하였다.

Lee와 Nelder (1996)은 다단계 일반화 선형모형의 여러 종류의 성분을 검정하기 위해 다양한 이탈도(deviance)들을 정의하였다. 변량효과를 검정하기 위해서는 이탈도로 $-2h$를, 고정효과를 검정하기 위해서는 주변 로그 가능도 l을 이용한 $-2l$을, 그리고 산포모수에 대해서는 $-2\log f_\theta(y|\hat{\beta})$를 사용할 것을 제안했다. 주변로그 가능도 l과 $\log f_\theta(y|\hat{\beta})$을 수치적으로 구하기 어려운 경우, 해당되는 근사 함수로 $p_v(h)$와 $p_{\beta,v}(h)$를 사용한다. 실제 검정방법은 Lee 등 (2006)에 나와 있다.

3.3. 이중 다단계 일반화 선형모형

지금까지 살펴본, 다단계 일반화 선형모형은 평균 모형에 변량효과를 포함 시킨 모형이다. 평균 모형 외에도 분산모형에도 변량 효과를 포함시켜 모형을 확장할 수 있다. Lee와 Nelder (2006)는 평균과 분산에 모두 변량효과를 도입한 이중 다단계 일반화 선형모형(Double hierarchcial generalized linear models; DHGLMs)을 개발하였다. 평균뿐만 아니라 분산에 있어서 개체들 또는 군집들 사이의 이질성이 모형에 고려되어야 할 때, 사용될 수 있는 모형이다. 뿐만 아니라, 경제학에서 사용되는 ARCH(autoregressive conditional heteroskedasticity; Engel, 1982)나 GARCH(generalized ARCH) 등의 모형들을 포함하며, 오차가 정규성을 벗어난 이상치를 가지는 두꺼운 꼬리를 가지는 분포들에 대한 추론 등에 필요한 모형도 쉽게 확장이 가능하다. 모형을 추론함에 있어, 이중 다단계 일반화 선형모형은 내부적으로 여러개의 일반화 선형모형들이 결합한 것으로 생각할 수 있고, 앞에서 소개한 다단계 가능도를 이용한 일반적인 추론 방법을 그대로 쓰면 되기 때문에 통계적, 수치적으로 효율적인 알고리즘을 제공한다.

이중 다단계 일반화 선형모형의 기본적인 구조는 다음과 같다.

(i) 변량효과 쌍 (a,u)가 주어졌을 때, 반응변수 y는 다음과 같은 조건부 평균과 분산을 가진다.

$$E(y|a,u) = \mu \quad \text{and} \quad \text{var}(y|a,u) = \phi V(\mu).$$

다단계 일반화선형모형과 마찬가지로 평균 μ에 대한 선형 예측치는

$$\eta = g(\mu) = X\beta + Zv \tag{3.6}$$

로 생각할 수 있다. 이때, $v = v(u)$이다.

이중 다단계 일반화 선형모형으로의 확장에 있어 중심이 되는 요소는 산포 모수 ϕ에 대하여 변량효과를 도입하는 것이다. 연결함수 $h(\ )$를 통하여, ϕ에 대하여 또 하나의 일반화 선형모형을 모형화 한다.

$$\xi = h(\phi) = G\gamma + Fb \tag{3.7}$$

이 때, G와 F는 고정효과 γ와 변량효과 $b = b(a)$에 해당하는 모형행렬을 의미한다.

(ii) 변량효과 u는 산포모수 λ를, 변량효과 a는 산포모수 α를 가지는 지수족 분포를 따른다.

이 모형에서 산포모수는 ϕ 외에도 평균 모형에 쓰인 변량효과 u의 산포모수 λ, 분산 모형에 쓰인 변량효과 a의 산포모수 α가 더 있다. 실제로 이중 다단계 일반화선형모형은 λ와 α 각각에 대해서도 일반화 선형모형을 적용할 수 있다 (Lee 등, 2006).

이중 다단계 일반화 선형모형에서의 추론을 위해서는 다음과 같이 다단계 가능도를 정의할 수 있다.

$$h = \log f(y|v, b; \beta, \phi) + \log f(v; \lambda) + \log f(b; \alpha),$$

여기서 $f(y|v, b; \beta, \phi)$, $f(v; \lambda)$와 $f(b; \alpha)$는 각각 (v, b)가 주어졌을 때 y의 조건부 밀도함수, v와 b의 밀도함수를 나타낸다. 주변 가능도 $L_{v,b}$는 h를 변량효과 v, b에 대하여 적분하여 다음과 같이 얻을 수 있으며

$$L_{v,b} = \log \int \exp(h) dv db = \log \int \exp L_v db = \log \int \exp L_b dv,$$

$L_v = \log \int \exp(h) dv$, $L_b = \log \int \exp(h) db$이다. 다단계 가능도 h와 주변 가능도 $L_v, L_{b,v}$는 변량효과와 고정효과들에 대한 적절한 추론을 제공한다. L_v와 $L_{b,v}$계산이 수치적으로 어려울 때는, 앞에서 설명한바와 같이 $p_v(h)$ $p_{v,b}(h)$ 등을 사용하는 것을 권장한다.

3.4. 다변량 이중 다단계 일반화 선형모형

반응변수들이 다변량으로 관측이 되고, 각각의 반응변수 간의 상관관계를 모형에 반영하기 위해서는 서로 상관된(correalted) 변량 효과들을 모형에 사용하면 된다. 먼저, 반응변수별로 적절한 이중 다단계일반화 선형모형(DHGLMs)을 만들고, 이들 모형에 들어있는 변량효과 간에 상관성을 고려함으로써, 서로 결합시킨 모형을 다변량 이중 다단계 일반화선형모형(multivariate DHGLMs; MDHGLMs)이라고 부른다. MDHGLMs을 사용하게 되면, 여러 형태(연속형, 이산형 등)를 갖는 반응변수들의 상관관계를 혼합 모형에 반영하여 분석 할 수 있다는 장점이 있다 (Lee와 Noh, 2012; Molas 등, 2013).

MDHGLMs의 예시로 (Price 등, 1985)의 ethylene glycol 독성 실험 자료를 살펴보자. 이 실험의 반응 변수는 두 가지로 하나는 쥐 태아의 몸무게(연속형 반응변수)와 기형 여부(이산형 반응변수)이며 약물의 투여 정도에 따라 쥐 태아의 기형 여부와 몸무게의 변화 여부를 살펴보기 위한 실험이다. 기형여부와 몸무게 사이에 상관성이 존재할 것이라고 생각된다면, 다음과 같은 MDHGLMs을 수립하여, 쉽게 모형화 할 수 있다. i번째 어미쥐의 j번째 태아의 몸무게와 기형여부를 y_{1ij}, y_{2ij}라 한다면, 변량효과 w_i, u_i가 주어졌을 때 다음과 같은 조건부 모형을 고려한다.

$$\begin{aligned} y_{1ij}|w_i &\sim N(\mu_{ij}, \phi), \quad \mu_{ij} = x_{1ij}\beta_1 + w_i, \\ y_{2ij}|u_i &\sim \text{Bernoulli}(p_i), \quad \text{logit}(p_{ij}) = x_{2ij}\beta_2 + u_i. \end{aligned}$$

만약 쥐 태아의 몸무게와 기형여부가 독립적이라고 가정한다면, 각각 별개의 HGLM 모형을 적합하면 된다. 하지만, 몸무게와 기형여부간의 상관성을 고려하기 위해, 두 변량효과 w_i와 u_i 간에 식 (3.8)과 같이 이변량 정규분포를 고려하면, 두 HGLM 모형은 결합된 형태로 하나의 MDHGLM 모형으로 생각할 수 있다.

$$\begin{pmatrix} w_i \\ u_i \end{pmatrix} \sim N\left(0, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right). \tag{3.8}$$

다단계 가능도를 이용하여 각각의 모수들을 추정하게 되면 약물의 효과(β)와 더불어 기형여부와 몸무게의 상관관계(ρ)를 동시에 추정하고 검정할 수 있다.

4. 다단계 일반화 선형모형의 활용

4.1. 평활법

평활법(smoothing)은 비모수적 함수 추정에서 많이 쓰이는 방법으로 평균 모형을 선형으로 적합 하는 대신 임의의 매끄러운 함수 $f(x)$를 이용하여

$$E(y|x) = f(x)$$

를 만족시키는 적합 방법이다. $f(x)$의 매끄러운 정도를 조절하기 위해 가장 보편적으로 사용되어 지는 방법은 벌점최소제곱(penalized least square) 방법으로 아래의 식을 최소화 하는 $f(x)$를 구한다.

$$\sum_i (y_i - f(x_i))^2 + \rho \int \left| f^{(d)}(x) \right|^2 dx, \tag{4.1}$$

여기서 $f^{(d)}(x)$는 $f(x)$의 d차 미분이고, ρ는 평활모수이다. $f(x)$를 q개의 기저 함수($B_j(x)$)를 이용하여 일반적인 스플라인 모형으로 만들면

$$f(x) = \sum_{j=1}^{q} v_j B_j(x)$$

로 표현 할 수 있으며 이때 회귀분석 모형의 형태로 다시 써보면

$$y_i = \sum_{j=1}^{q} v_j B_j(x_i) + e_i$$

이고, 행렬 형태로는

$$y = Zv + e$$

이다. 여기서 모형 행렬 Z의 원소는 $z_{ij} \equiv B_j(x_i)$로 정해진다. 이렇게 나타낼 경우 식 (4.1)에서 벌칙함수는

$$\rho \int \left| f^{(d)}(x) \right|^2 dx \equiv \rho v^t P v,$$

로 나타낼 수 있고 여기서 P의 (i,j)번째 원소는

$$\int B_i^{(d)}(x) B_j^{(d)}(x) dx$$

로 표현할 수 있다. 즉, 결국 식 (4.1)은

$$||y - Zv||^2 + \rho \sum_j |\Delta^2 v_j|^2 \equiv ||y - Zv||^2 + \rho v^t P v$$

로 선형 혼합모형의 형태로 표현이 된다. 결과적으로 평활법 또한 변량효과모형으로 설명 할 수 있으며 기존의 선형혼합 모형 또는 다단계 일반화 선형모형의 추정 방법을 통해 함수추정이 가능하다. 이것에 대한 좀 더 구체적인 설명은 Lee 등 (2006)을 참고하기 바란다.

4.2. 생존자료 분석에서의 변량효과

생존자료 분석에서도 변량효과의 적용이 가능하다. 생존 자료 분석에 사용되는 대표적인 모형인 Cox의 비례 위험 모형에 변량효과(즉 프레일티)를 포함한 모형인 프레일티 모형(frailty model)이 그 대표적인 예이다. 특히 프레일티 모형은 개인이나 군(cluster)에 의해 생존자료가 얻어지는 다변량생존자료(correlated survival data)의 분석에 매우 유용하다.

$T_{ij}\,(i = 1, \ldots, q,\ j = 1, \ldots, n_i,\ n = \sum_i n_i)$를 i번째 사람의 j번째 생존시간이라 하고 C_{ij}를 대응하는 중도 절단시간이라고 하자. 그러면 관측가능한 확률변수는 $Y_{ij} = \min(T_{ij}, C_{ij})$와 중도절단 여부인 $\delta_{ij} = I(T_{ij} \le C_{ij})$이다. u_i를 i번째 사람의 변량효과로 하고, u_i가 주어졌을 때 조건부 T_{ij}의 위험 함수는

$$\lambda_{ij}(t|u_i) = \lambda_0(t) \exp\left(x_{ij}^T \beta\right) u_i$$

으로 모형화한다. 여기서 $\lambda_0(\cdot)$는 모수적 또는 비모수적 기저(baseline) 위험함수이다. 이때 u_i의 분포는 ξ를 모수로 갖는 분포로 정의되며 주로 감마분포나 로그정규분포를 사용한다.

$v_i = \log(u_i)$로 두고 f_{2i}를 v_i의 확률밀도함수라 할 때, 프레일티 모형에 대한 Lee와 Nelder (1996)의 다단계 가능도는 다음과 같이 표현할 수 있다 (Ha 등, 2001):

$$h = h(\beta, \Lambda_0, \xi) = \sum_{ij} \ell_{1ij} + \sum_i \ell_{2i},$$

여기서 $\ell_{1ij} = \ell_{1ij}(\beta, \Lambda_0; y_{ij}, \delta_{ij}|u_i) = \log f_{1ij}$, $\ell_{2i} = \ell_{2i}(\xi; v_i) = \log f_{2i}$이고,

$$f_{1ij} = \{\lambda(y_{ij}|u_i)\}^{\delta_{ij}} \exp\{-\Lambda(y_{ij}|u_i)\}$$

로 표현된다. 여기서 $\Lambda(\cdot)$는 누적위험함수이다. 이에 대한 보다 상세한 추론방법 및 예제는 Ha 등 (2012)을 참고하기 바란다.

4.3. 결측치 처리

일반적인 가능도 이론은 모든 관측치가 갖추어진(complete) 자료를 기본으로 한다. 그러나 자료에는 관측되지 않거나 누락된 자료가 존재하는 경우가 종종 있다. 이러한 문제를 해결하기 위해 많은 연구들이 진행 되어 왔으며 변량효과를 이용한 결측자료(missing data) 분석 또한 가능하다. 결측없이 모두 관측했다고 가정할 때, $y_i^* = (y_{i1}, \ldots, y_{iJ})$를 i번째 사람의 모든 자료라 하고, y_i^* 중에 실제로 관측한 값들을 y_i^O, 결측에 해당하는 부분을 y_i^M이라고 하자. 또한 $R_i = (R_{i1}, \ldots, R_{iJ})$를 결측 여부를 알려주는 벡터라고 하자. 즉, j번째 관측치가 결측이면 $R_{ij} = 0$이고, 관측이 되면 $R_{ij} = 1$이다.

y_i^*와 R_i의 결함분포 함수는 아래와 같이 계산할 수 있으며

$$f\left(y_i^*, R_i; \theta, \lambda\right) \equiv f\left(y_i^*; \theta\right) f\left(R_i | y_i^*; \lambda\right). \tag{4.2}$$

이때, 결측값 y_i^M를 관측하지 못한 변량 v_i로 처리하고, 다단계 가능도를

$$h = \sum_i f\left(y_i^*, R_i; \theta, \lambda\right) = \sum_i \log f\left(y_i^O, v_i, R_i; \theta, \lambda\right)$$

로 세움으로써, 기존의 다단계 가능도 이론으로 혼합모형 안에서 고정된 모수 (θ, λ)와 결측값인 변량 $v_i = y_i^M$에 대한 추론이 동시에 가능해진다.

4.4. 변수선택법

모형에 사용되는 설명 변수의 수가 매우 많은 경우, 모든 변수들이 실제로 유의미한 경우는 드물다. 그래서 유의미한 변수들을 선택하기 위해 벌칙 가능도(penalized likelihood; PL)방법이 많이 연구 되어졌다. 벌칙 가능도 방법은 기존의 가능도에 적절한 벌칙함수(penalty function)를 붙임으로써 벌칙 가능도를 최대화 하는 모수를 추정하다보면 자동적으로 의미 있는 변수가 선택되는 방법이다. 먼저 간단한 선형 모형을 가정하여 보자.

$$\begin{aligned} y_i &= x_i^T \beta + \epsilon_i \quad i = 1, \ldots, n \\ \epsilon_i &\overset{iid.}{\sim} N(0, \phi). \end{aligned} \tag{4.3}$$

이때 다음과 같은 $Q_\lambda(\beta)$를 최소로 하는 회귀계수 β를 찾는 방법이 벌칙가능도 방법이다.

$$Q_\lambda(\beta) = \frac{1}{2} \sum_{i=1}^{n} \left(y_i - x_i^T \beta\right)^2 + \sum_{j=1}^{d} p_\lambda(|\beta_j|),$$

여기서 $p_\lambda(|\beta_j|)$는 주어진 벌칙함수이고, λ는 벌칙함수의 모양을 결정해주는 모수이다. Lee와 Oh (2014)는 회귀계수 β를 변량효과로 보고 혼합효과모형을 적합하여 벌칙가능도 방법에서와 같은 해를 구하고, 이론적으로 좋은 성질을 가진 새로운 벌칙함수를 제안하였다. Lee와 Oh (2014)는 또 하나의 변량효과 u가 주어졌을 때, β가 정규분포를 따른다고 가정하였다:

$$\beta | u \sim N(0, u\theta).$$

그리고 정규 분포의 분산성분에 들어있는 변량 u는 $E(u) = 1$, $\mathrm{Var}(u) = w$를 가지는 감마 분포를 따른다고 가정한다. 이러한 변량효과 모형을 가정하고 다단계 가능도를 이용하여 Lee와 Oh (2014)는 식 (4.3)에서 벌칙함수 $p_\lambda(|\beta_j|)$이 다음과 같이 표현됨을 보였다.

$$\begin{aligned} p_\lambda(|\beta_j|) &= \frac{\phi}{2\theta} \frac{\beta_j}{\hat{u}} + \frac{\phi(w-2)}{w} \log \hat{u} + \frac{\phi}{w} \hat{u}, \\ \hat{u} &= \hat{u}(\beta_j) = w \left\{ \left(\frac{2}{w} - 1\right) + \sqrt{\frac{8\beta_j^2}{w\theta} + \left(\frac{2}{w} - 1\right)^2} \right\} \Big/ 4. \end{aligned}$$

이러한 방법으로 제안된 벌칙함수는 w의 값에 따라 그 모양이 변화하게 되며, 특히 w가 2보다 큰 경우 0에서의 값이 음의 무한대값을 갖는 형태가 된다. 이러한 새로운 벌칙함수는 통계학의 전반적인 부분에 사용되고 있으며 그 이론적 성질 또한 여러 논문에서 다뤄지고 있다. 특히, 선형 모형, 일반화선형

모형의 일반적인 회귀모형 외에도, 주성분분석, 정준 상관분석, 부분 최소제곱법과 같은 다변량 자료 분석 및 프레일티 모형, 경쟁 위험 모형과 같은 생존 자료 분석에 많이 응용되어 변수선택에 있어서 탁월한 성능을 보인바 있다 (Lee와 Oh, 2014; Lee 등, 2010, 2011; Ha 등, 2014). 또한 변수 간의 그룹정보가 있는 경우, 그룹정보를 반영하여 변수선택을 하는 방법 또한 개발 되었다 (Lee 등, 2015).

4.5. 다중검정법

최근 기술의 발전에 의해 유전학이나 영상자료 분석에서는 수천개의 가설을 동시에 검정해야 하는 일이 벌어진다. 이러한 다중검정 문제를 해결하기 위한 방법으로 p-value를 이용한 Benjamini와 Hochberg (1995)의 오발견율(False discovery rate; FDR) 제어 방법 등이 제안되었고, 많이 사용되고 있다. 2.2절에서 언급하였듯이, 다중검정 문제 또한 관측하지 못한 변량이 모형에 들어있어야 하므로, Lee와 Bjørnstad (2013)는 다중 검정 문제를 관측하지 못한 변량의 예측 문제로 보고, 확장된 가능도를 이용한 변량효과 모형과 추론을 통해 효율적으로 오발견율을 제어하기 위한 방법을 제안하였다. 특히, Lee와 Bjørnstad (2013)는 p-value와 같은 요약통계량을 기초로한 검정 방법은 비효율적임을 지적하고, 최적의 검정방법은 올바른 통계모형과 확장된 가능도를 통해서 해결 될 수 있음을 보였다. y_{ij}를 i번째 가설 H_i에 해당하는 j번째 개체의 자료라고 하자 ($i = 1, \ldots, N$, $j = 1, \ldots, n$,). 여기서, y_{ij}의 모형으로 효과크기 w_i와 오차항 e_{ij}를 가지는 선형 모형

$$y_{ij} = w_i + e_{ij}, \quad E(e_{ij}) = 0,\ \mathrm{Var}(e_{ij}) = \phi_i$$

을 가정한다. 각 H_i별로, 이진형 변량 o_i를 도입한다. 즉, $o_i = 1$이면 i번째 귀무가설이 거짓인 경우이며, $o_i = 0$은 i번째 귀무가설이 참인 경우라 하자. 귀무가설을 효과크기 w_i가 거의 0에 가까움이라고 할 때, o_i에 따라 다음처럼 w_i의 평균과 분산을 가정할 수 있다.

$$\begin{aligned} o_i = 0\text{일때}, &\quad E(w_i) = 0,\ \mathrm{Var}(w_i) = \sigma^2, \\ o_i = 1\text{일때}, &\quad E(w_i) = \mu,\ \mathrm{Var}(w_i) = \tau^2. \end{aligned}$$

이때, 식을 정리하면, 통계량

$$t_i = \frac{\bar{y}_i}{\sqrt{\phi_i/n + \sigma^2}}$$

가 따르는 분포는 아래와 같다.

$$\begin{aligned} o_i = 0\text{일때}, &\quad t_i \sim f_0,\ \text{iid}\ \ E(t_i) = 0,\ \mathrm{Var}(t_i) = 1, \\ o_i = 1\text{일때}, &\quad t_i \sim f_{1i},\ \text{independent}\ E(t_i) = \mu_i^*,\ \mathrm{Var}(t_i) = \varphi_i, \\ &\qquad \mu_i^* = \frac{\mu}{\sqrt{\phi_i/n + \sigma^2}},\ \varphi_i = \frac{\left(\phi_i/n + \tau^2\right)}{\left(\phi_i/n + \sigma^2\right)}. \end{aligned}$$

위의 모델에서와 같이 귀무가설의 참 거짓 유무를 관측되지 않은 이진형 변량효과로 놓음으로서 확장된 가능도 방식으로 다중검정 문제를 이진형 변량효과 o_i를 예측하는 문제로 생각할 수 있다. 실제 관련 모수 및 오발견율을 추정하고, 주어진 손실함수를 최소화 하는 최적화된 다중검정 방법은 Lee와 Bjørnstad (2013)를 참고하길 바란다. 최근에는 Lee와 Bjørnstad (2013) 방법을 확장하여 o_i가 이진형이 아닌 여러개의 카테고리를 가지고 있는 이산형일 때의 다중 검정 문제나, 각 검정 간 상관관계가 있는 다중검정의 문제에 대한 연구도 진행되고 있다.

5. 결론

기존의 통계학 모형들은 관측 가능한 변량들 사이의 관계만을 설명하였다. 따라서, 관측할 수 없는 변량효과를 이용하여 관측된 변량들을 설명하는 새로운 통계학 모형들이 대두되었다. 그러나, 기존의 Fisher의 가능도는 관측할 수 없는 변량들에 대한 추론을 할 수 없으므로 다단계 가능도로 확장하여야 한다. 이는 통계학의 새로운 패러다임으로서 이를 이용하여 변수들 간의 상관, 미래 예측, 분류 등 여러 다양한 분야에서 사용될 수 있다. 한편, 이런 모형들을 이용하여 인과관계의 규명, 또는 구조방정식과의 융합을 통해 기존 통계학의 범위를 넓히는 새로운 길을 열어야 한다. 이 논문에서는 관측되지 않는 변량이 포함된 혼합효과모형을 위주로 논하고, 다단계 일반화 선형모형과 그 확장된 모형을 중심으로 간단히 소개를 하였다. 장래에 통계적 문제들이 점점 다양해지고 복잡해짐에 따라 변량효과를 포함하는 복잡한 모형들과 이를 분석하기 위한 다단계 가능도에 대한 연구와 자료 분석 등의 수요는 점점 더 커질 것이다. 한국의 통계학자들도 혼합효과모형과 확장된 가능도 이론의 발전에 중요한 역할을 하기를 바라면서 글을 마친다.

References

Airy, G. B. (1861). *On the Algebraic and Numerical Theory of Errors of Observations and the Combination of Observations*, Macmillan and Co., Ltd., London.

Bayarri, M. J., DeGroot, M. H. and Kadane, J. B. (1988). What is the likelihood function? (with discussion), *Statistical Decision Theory and Related Topics IV*, **1**, eds S.S. Gupta and J. O. Berger, Springer, New York.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. by the late Rev. Mr. Bayes, FRS. communicated by Mr. Price, in a letter to John Canton, AMFRS, *Philosophical Transactions (1683-1775)*, 370–418.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society B*, **57**, 289–300.

Berger, J. O. and Wolpert, R. (1984). The Likelihood Principle, *Hayward: Institute of Mathematical Statistics Monograph Series*.

Birnbaum, A. (1962). On the foundations of statistical inference, *Journal of the American Statistical Association*, **57**, 269–326

Bjørnstad, J. F. (1996). On the generalization of the likelihood function and likelihood principle, *Journal of the American Statistical Association*, **91**, 791–806.

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

Butler, R. W. (1986). Predictive likelihood inference with applications (with discussion), *Journal of the RoyalStatistical Society, Series B*, **48**, 1–38.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation, *Econometrica*, **50**, 987–1008.

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample, *Metron*, **1**, 3–32.

Ha, I. D., Lee, Y. and Song J.-K. (2001). Hierarchical likelihood approach for frailty models, *Biometrika*, **88**, 233–243.

Ha, I. D., Noh, M. and Lee, Y. (2012). frailtyHL: A package for fitting frailty models with h-likelihood, *R Journal*, **4**, 28–37.

Ha, I. D., Pan, J., Oh, S. and Lee, Y. (2014). Variable selection in general frailty models using penalized h-likelihood, *Journal of Computational and Graphical Statistics*, **23**, 1044–1060.

Lauritzen, S. L. (1974). Sufficiency, prediction and extreme models, *Scandinavian Journal of Statistics*, **1**, 128–134.

Lee, D., Lee, W., Lee, Y. and Pawitan, Y. (2010). Super-sparse principal component analyses for high-

throughput genomic data, *BMC Bioinformatics*, **11**, 296.

Lee, D., Lee, W., Lee, Y. and Pawitan, Y. (2011). Sparse partial least-squares regression and its applications to high-throughput data analysis, *Chemometrics and Intelligent Laboratory Systems*, **109**, 1–8

Lee, J., Lee, K. and Lee, Y. (2014). History and future of Bayesian statistics, *The Korean Journal of Applied Statistics*, **27**, 855–863.

Lee, S., Pawitan, Y. and Lee, Y. (2015). A random-effect model approach for group variable selection, *Computational Statistics and Data Analysis*, **89**, 147–157.

Lee, Y. and Bjørnstad, J. F. (2013). Extended likelihood approach to large scale multiple testing, *Journal of the Royal Statistical Society B*, **75**, 553–575.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion), *Journal of the Royal Statistical Society B*, **58**, 619–678.

Lee, Y. and Nelder, J. A. (2001). Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions, *Biometrika*, **88**, 987–1006.

Lee, Y. and Nelder, J. A. (2006). Double hierarchical generalized linear models (with discussion), *Journal of the Royal Statistical Society C*, **55**, 139–185.

Lee, Y., Nelder, J. A. and Pawitan, Y. (2006). *Generalised Linear Models with Random Effects: Unified Analysis via h-Likelihood*, Chapman and Hall, London.

Lee, Y. and Noh, M. (2012). Modelling random effect variance with double hierarchical generalized linear models, *Statistical Modelling*, **12**, 487–502.

Lee, Y. and Oh, H. (2014). A new sparse variable selection via random-effect model, *Journal of Multivariate Analysis*, **125**, 89–99.

Molas, M., Noh, M., Lee, Y. and Lesaffre, E. (2013). Joint hierarchical generalized linear models with multivariate Gaussian random effects, *Computational Statistics and Data Analysis*, **68**, 239–250.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models, *Journal of the Royal Statistical Society A*, **135**, 370–384.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference using Likelihood*, Clarendon Press, Oxford.

Pearson, K. (1920). The fundamental problems of practical statistics, *Biometrika*, **13**, 1–16.

Price, C. J., Kimmel, C. A., Tyle, R. W. and Marr, M. C. (1985). The developmental toxicity of Ethylene Glycol in rates and mice, *Toxicological Applications in Pharmacololgy*, **81**, 113–127.

혼합효과모형의 리뷰

이영조[a,1]

[a]서울대학교 통계학과

(2015년 4월 16일 접수, 2015년 4월 23일 수정, 2015년 4월 23일 채택)

요 약

관측 가능한 변수들 사이의 관계를 묘사한 갈릴레오의 물리학 법칙 발견 이후, 과학은 큰 성과를 거두며 발전해왔다. 그러나, 관측할 수 없는 변량효과를 함께 이용하여 더 많은 자연 현상을 설명할 수 있게 되었고, 이를 이용한 최초의 통계적 모형인 혼합효과모형이 소개되었다. 계산기술의 발달과 더불어 복잡한 현상에 대한 추론을 위하여 혼합효과모형은 그 중요성이 더욱 커지고 있다. 이러한 혼합효과모형은 최근 다단계 일반화 선형모형을 포함한 여러 모형으로 확장되었으며, 관측할 수 없는 변량효과를 추론하기 위한 다단계 가능도가 제시되었다. 혼합효과모형 특집호를 통해 이러한 모형들이 여러 통계학적 문제점을 해결하는 과정을 제시하고, 앞으로 어떤 확장이 추가적으로 요구되는지에 대하여 논할 것이다. 빈도록적 접근법과 베이지안 접근법을 함께 다룬다.

주요용어: 혼합효과모형, 다단계 일반화 선형모형, 다단계 가능도, 변량 효과.

이 논문은 한국연구재단의 지원을 받아 수행된 기초연구사업(No. 2011-0030810)과 미래창조과학부의 뇌과학원천기술개발사업(2014M3C7A1062896)으로부터 지원받아 수행되었습니다.

[1](151-742) 서울특별시 관악구 관악로 1, 서울대학교 통계학과. E-mail: youngjo@snu.ac.kr