The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.2.335
28(2), 335–342

The Use of Joint Hierarchical Generalized Linear Models: Application to Multivariate Longitudinal Data

Donghwan Lee[a,1] · Jae Keun Yoo[a]

[a]Department of Statistics, Ewha Womans University

(Received March 24, 2015; Revised March 31, 2015; Accepted March 31, 2015)

Abstract

Joint hierarchical generalized linear models proposed by Molas *et al.* (2013) extend the simple longitudinal model into multiple models fitted jointly. It can easily handle the correlation of multivariate longitudinal data. In this paper, we apply this method to analyze KoGES cohort dataset. Fixed unknown parameters, random effects and variance components are estimated based on a standard framework of h-likelihood theory. Furthermore, based on the conditional Akaike information criterion the correlated covariance structure of random-effect model is selected rather than an independent structure.

Keywords: Joint hierarchical generalized linear models, mixed models, cohort data, longitudinal data.

1. 서론

경시적 자료(longitudinal data)는 일정 기간 동안 각 개체(subject)마다 시간에 따라 반복 측정되는 자료로써, 코호트 연구 등에서 많이 쓰인다. 따라서, 반복 측정된 개체에서 나온 관측치들은 서로 상관관계가 있고, 이러한 상관관계를 고려하기 위해 다양한 혼합 모형들이 제안되었다. 특히, 반응변수들의 정규성을 가정하기 어려운 범주형 자료분석인 경우에는 일반화 선형 혼합 모형(generalized linear mixed models; GLMMs, Breslow와 Clayton, 1993), 또한 이를 포함하는 다단계 일반화 선형 혼합 모형(hierarchical generalized linear models; HGLMs, Lee와 Nelder, 1996) 등이 많이 쓰인다. Lee와 Nelder (1996)은 관측되지 않는 변량 효과(random effects)를 도입하여 반복측정 된 관측치들의 상관관계를 쉽게 설명할 수 있음을 보였다. 하지만, 경시적 자료의 수집에 있어서 동일한 개체에서 2개 이상의 반응변수들이 관측이 되고, 이들 간에도 상관관계가 존재한다면, 다변량 경시적 자료를 다룰 수 있는 모형의 확장이 필요하다. 경시적 자료에서 다변량 반응변수 사이의 상관성은 공분산 구조를 가지는 변량효과를 고려하면 된다 (Lee와 Nelder, 2001). Yun과 Lee (2004)는 이변량 이진-정규 혼합 모형을 다단계 우도를 이용하여 적합하였다.

Lee 등 (2009)은 주변화 된 변량효과 모형(marginalized random effect models)을 이용하여 다변량 경시적 이진 자료 분석을 위한 방법론을 제안하였다. 자료 간 상관성을 고려하기 위해 변량효과를 사용

This research was supported by the Brain Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (2014M3C7A1062896).
[1]Corresponding author: Department of Statistics, Ewha Womans University, Seoul 120-750, Korea.
E-mail: donghwan.lee@ewha.ac.kr

하면서도, 주변 모형에 초점을 맞춤으로써 개체별 조건부 회귀 계수 추정 보다 모집단 평균 회귀 계수 추정이 주 관심사인 경우에 적합한 분석 방법이다. 최근, Molas 등 (2013)은 각 반응변수들이 HGLM 모형을 따르고, 변량효과들을 여러가지 상관구조를 가질 수 있는 일반적인 다변량 정규분포를 따른다고 가정하여, 반응변수 간에 상관성을 고려한 결합 다단계 일반화 선형모형(joint hierarchical generalized linear models; JHGLMs)을 제안하였다. 이 모형의 장점은 각 다변량 반응변수들이 같은 유형이어야 하는 제약 조건이 필요 없다. 즉, 한 반응변수는 정규분포를, 다른 반응변수는 이항분포를 따른다고 가정하는 것이 가능하다. 또한 Molas 등 (2013)은 R 패키지 `mdhglm` 제공하여, 변량 효과들이 상관관계 있을 때 이외에도 서로 독립(independent)일 때, 포화 관계(saturated)일 때의 모형도 제공함으로써 가장 적합한 혼합 모형을 선택하는데 효과적이다.

본 논문의 구성은 다음과 같다. 2장에서는 다변량 경시적 자료를 다루기 위한 결합 다단계 일반화 선형모형과 모수 추정 방법을 살펴보고, 3장에서는 실제 자료인 Korean Genomic Epidemilogy Study(KoGES)에서 수행한 코호트 자료를 이용하여 분석을 하고 가장 적합한 모형을 선택해본다. 마지막으로 4장에서는 결론을 제시한다.

2. 방법론

이 장에서는 Molas 등 (2013)이 제안한 결합 다단계 일반화 선형모형을 소개하고 관련 모수 추정 방법을 요약한다.

2.1. 결합 다단계 일반화 선형모형

다변량 경시적 자료에서 i번째 개체(subject) ($i = 1, \ldots, N$)의 j번째 시간(time) ($j = 1, \ldots, n_i$)에 관측된 K-변량 반응변수 벡터 $y_{ij} = (y_{ij1}, \ldots, y_{ijK})^T$라고 하자. 이 때, 변량효과 $v = (v_{11}, \ldots, v_{1K}, v_{21}, \ldots, v_{2K}, \ldots, v_{N1}, \ldots, v_{NK})^T$가 주어졌을 때, 종속변수 각각의 반응변수 y_{ijk}는 조건부 평균 μ_{ijk}를 갖는 어느 한 지수족(exponential family)의 분포를 따르고, 다음을 만족한다.

$$g_k(\mu_{ijk}) = x_{ijk}^T \beta_k + v_{ik}, \tag{2.1}$$

여기서 $g_k(\ )$는 연결함수(link function), x_{ijk}는 설명변수(explanatory), β_k은 고정효과(fixed parameter)를 나타낸다.

변량효과 벡터 v는 평균이 0이고 공분산 행렬이 Σ인 다변량 정규분포를 따른다고 가정한다. 이 때, 공분산 행렬 Σ를 구성하는 분산성분 모수 벡터를 τ라고 한다 ($\Sigma = \Sigma(\tau)$). 예를 들어, $K = 2$이고, 개체 간 독립을 가정하고 개체 내 상관성을 고려하는 경우에

$$\begin{aligned} \Sigma &= I_N \otimes D, \\ D &= \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \end{aligned} \tag{2.2}$$

이고, 분산 성분 모수 벡터 $\tau = (\sigma_1, \sigma_2, \rho)^T$가 된다. $\rho = 0$인 경우에는 모든 변량효과가 독립임을 가정한다.

2.2. 모수 추정

모형을 적합하는데 필요한 모수 및 변량효과들을 추정하기 위해, Lee와 Nelder (1996)으로부터 다음과

같은 다단계 우도(hierarchical likelihood)를 정의한다.

$$h = \sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{k=1}^{K} \log f_\phi(y_{ij}|v) + \log f(v), \tag{2.3}$$

여기서 $\log f_\phi(y_{ijk}|v)$는 분산모수 ϕ를 가지는 y_{ij}의 로그 조건부 확률 밀도 함수를 의미하고,

$$\log f(v) = -\frac{NK}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}v^T\Sigma^{-1}v \tag{2.4}$$

이다. 위의 모형 (2.1), (2.4)에서 추정해야할 대상은 변량효과 v, 평균 모수 벡터 $\beta = (\beta_1^T, \ldots, \beta_K^T)^T$와 분산 모수 $\theta = (\phi, \sigma_1, \sigma_2, \rho)$이다.

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

$$p_\alpha(l) = \left[l - \frac{1}{2}\log\left|\frac{D(l,\alpha)}{2\pi}\right|\right]\Bigg|_{\alpha=\tilde{\alpha}} \tag{2.5}$$

여기서 l은 로그 주변 우도 또는 다단계 우도를 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 (2.3)의 h를 최대화 시키는 값을, 평균 모수인 β는 수정된 단면우도 $p_v(h)$를, 분산 성분인 θ는 $p_{v,\beta}(h)$를 최대화 시키는 값으로 구하였다. 본 연구에서는 필요한 모수들들 추정하기 위해 사용한 R패키지 `dhglm`와 보다 자세한 모수 추정 절차는 Molas 등 (2013)에 잘 나와 있다.

2.3. 모형 선택

내포 모형(nested model) 관계에 있는 두 모형을 비교할 때에는 우도비 검정(likelihood ratio test)를 이용하면 된다. 내포 여부에 상관없이, 변량효과모형을 비교하는 경우에는 조건부 아카이케 정보 기준(conditional Akaike information criterion; cAIC)을 이용하여 모형 선택을 할 수 있다. cAIC는 변량효과모형에서 같은 변량효과가 주어졌을 때, 반응변수와 독립인 미래 예측값을 이용하여 정의된 조건부 아카이케 정보(conditional Akaike information)의 점근적 불편추정량이므로, 변량효과모형의 최적 모형 선택에 효과적이고, 계산에 있어서도 필요한 양들이 모수 추정 시에 모두 계산됨으로 특별히 추가 계산이 필요 없는 장점이 있다 (Donohue 등, 2011). cAIC는 다음과 같이 정의된다:

$$\text{cAIC} = A + 2df.$$

여기서, A와 df는

$$A = -2h_1(\hat{\mu}; y|v) = \sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{k=1}^{K}\log f_\phi(y_{ij}|v)|_{\mu=\hat{\mu}}, \qquad df = \text{tr.}\left(H^{-1}H^*\right),$$

$$H = -\frac{2\partial^2 h}{\partial(\beta,v)\partial(\beta,v)^T}, \quad H^* = -\frac{2\partial^2 h_1}{\partial(\beta,v)\partial(\beta,v)^T}$$

로 구할 수 있다 (Lee 등, 2006).

3. KoGES 코호트 자료 분석

3.1. 자료 설명

KoGES에서 2001년부터 수행한 39세에서 69세사이의 중년성인 코호트 자료를 이용하였다. 코호트 자료의 자세한 설명은 Kim 등 (2006)을 참조한다. 코호트 참가자들은 2년에 한번 씩 대사 증후

군(metabolic syndrome)의 정도를 추적 관찰 받았다. Lee 등 (2009)은 여러 반응변수들 중에 고중성지방혈증(high hypertriglyceridemia; TG)과 고밀도 지방단백질 콜레스테롤(high-density lipoprotein cholestrol; HDL-C) 두 변수에 초점을 맞추었다. 또한, Lee 등 (2009)은 중도탈락(dropout) 여부가 임의가 아니라고 판단하여, pattern mixture model을 이용하여 중도탈락 메카니즘을 모형에 반영하였다. 이 논문에서는 분석의 단순화를 위해서 같은 중도탈락 메카니즘에 해당하는 처음 2회 방문 후 중독탈락을 하는 환자 207명만을 대상으로 하였다. 두 반응변수와 설명변수는 Lee 등 (2009)과 마찬가지로 TG값이 기준치(150mg/dl) 보다 높을 때 1의 값을 가지고, HDL-C는 기준치(남성: 40mg/dl, 여성: 50mg/dl) 보다 낮을 때 1의 값을 가지는 이진형 자료를 이용하였다. 설명변수로는 sex(남성 1; 여성 0), age(연령; 로그변환함), drink1(과거 음주여부), drink2(현재 음주여부), smok1(과거 흡연 여부), smoke2(현재 흡연여부)로 6개의 변수를 사용하였다.

3.2. 자료 분석

두 반응변수 TG와 HDL-C 모두 이진형(binary) 자료이기 때문에 반응변수들의 조건부 확률 분포로는 이항분포를, 연결함수는 로짓 링크를 사용한다. 초반 2회 이후에 중도탈락이 되는 자료들이기 때문에 모든 환자들의 반복 관찰수는 2회이다. 2장에서 설명한 모형을 자료 맞게 수정하면 다음과 같다: y_{ij1}, y_{ij2}을 i번째 환자의 j번째 관찰에서의 TG, HDL-C 값이라고 하면, 변량효과 v가 주어졌을 때 각각의 조건부 확률분포는

$$f(y_{ij1}|v) = \mu_{ij1}^{y_{ij1}}(1-\mu_{ij1})^{1-y_{ij1}},$$
$$f(y_{ij2}|v) = \mu_{ij2}^{y_{ij2}}(1-\mu_{ij2})^{1-y_{ij2}}$$

이고 조건부 평균 μ_{ij1}와 μ_{ij2}는

$$g_1(\mu_{ij1}) = \log\frac{\mu_{ij1}}{1-\mu_{ij1}} = x_{ij1}^T\beta_1 + v_{i1},$$
$$g_2(\mu_{ij2}) = \log\frac{\mu_{ij2}}{1-\mu_{ij2}} = x_{ij2}^T\beta_2 + v_{i2}$$

로 설명변수들과 연결된다. 각 i마다 이변량 변량효과 $v_i = (v_{i1}, v_{i2})^T$는 평균이 0이고 공분산이 (2.2)의 D와 같이 고려한다. 이 때, $\rho = 0$으로 고정시킨 모형을 독립모형(Model 1), ρ도 자료로부터 추정하는 모형을 상관모형(Model 2)라고 하여 각각의 모형들을 적합하고 비교한다. 또한, 성별에 따라서 고정효과와 변량효과들의 상관성이 다를 수도 있다고 판단하여, 전체 자료 적합 외에도 성별에 따라 각각 분석을 해보았다.

3.3. 결과

2개의 모형(독립, 상관)에 대하여 결합 다단계 일반화 선형모형을 적합시킨 후 고정 모수와 분산 모수 추정치들과 cAIC 값을 Table 3.1에 제시하였다. $\rho = 0$을 강제한 독립모형인 Model 1과 $\rho = 0$을 강제하지 않고 추정하는 Model 2 모두 반응변수가 TG인 모형에서 age의 효과가 양의 값으로 유의함을 알 수 있다. 즉, 연령이 올라갈수록 TG값이 기준치 이상이 될 가능성이 높아진다. 반응변수가 HDL-C인 모형에서는 Model 2의 경우 sex가, Model 1의 경우 drink2 변수가 유의하였다. 분산 모수들은 모두 유의하여, 관측치 간에 상관성이 존재하고, 변량효과가 모형에 필요함을 확인할 수 있다. 특히, Model 2에서의 ρ의 추정치가 양의 값으로 유의하고, cAIC값 또한 Model 2를 선택하는 것으로 보아 각 환자별 이 변량효과는 양의 상관관계에 있고, 두 반응변수 TG 수치와 HDL-C 수치의 증감에 서로 음의 영향을 미침을 확인 할 수 있다.

Table 3.1. Estimates of fixed parameters and variance components for joint hierarchical generalized linear models

	Model 1 (Independent)	Model 2 (Correlated)
Response variable: TG		
intercept	−20.235* (5.616)	−17.519* (4.416)
sex	0.770 (0.705)	0.502 (0.552)
age	4.776* (1.411)	4.169* (1.109)
drink1	−0.764 (0.763)	−0.774 (0.644)
drink2	0.225 (0.496)	0.177 (0.397)
smoke1	1.006 (0.848)	0.956 (0.682)
smoke2	1.106 (0.736)	1.003 (0.575)
Response variable: HDL-C		
intercept	−2.459 (5.322)	−4.153 (7.737)
sex	−1.137 (0.671)	−2.097* (0.944)
age	1.280 (1.343)	2.003 (1.958)
drink1	0.471 (0.813)	0.087 (1.073)
drink2	−0.950* (0.469)	−1.421 (0.619)
smoke1	−1.147 (0.817)	−0.871 (1.071)
smoke2	0.206 (0.694)	0.929 (0.948)
Variance components:		
$\log(\sigma_1^2)$	1.777* (0.224)	1.115* (0.210)
$\log(\sigma_2^2)$	1.543* (0.228)	2.609* (0.281)
ρ	-	0.649 (0.017)
$-2h$	2074.59	2044.669
$-2p_v(h)$	964.27	958.656
$-2p_{v,\beta}(h)$	954.67	947.465
cAIC	894.16	864.972

Standard errors are given in parentheses.

$*$: indicates significance with 95% Wald confidence interval.

다단계 우도 이론에 근거한 다단계 일반화 선형모형 및 결합 다단계 일반화 선형모형의 장점은 개체 각각의 특성을 설명해주는 변량효과의 추론이 다른 모수 추정들과 동시에 가능하다는 것이다. Figure 3.1에 Model 2의 변량효과 추정치들을 95% Wald 예측 구간을 함께 나타낸 것이다. $\hat{\rho} = 0.649 > 0$이기 때문에, 각각 환자 개개인의 HG와 HDL-C 모형에 들어가는 이변량 변량효과는 유사한 패턴을 보이는 것을 알 수 있다.

Table 3.2는 성별에 따라 그룹을 나누어서 각각에 대하여 JHGLMs를 적합시킨 결과이다. 남성 그룹의 경우 HDL-C 반응변수 모형에서 drink2(현재 음주여부) 변수가 유의하여, 현재 음주를 하게 되면, HDL-C가 낮아지는 것을 방해함을 알 수 있다. 여성 그룹의 경우에는 age(연령) 변수가 유의하여, 남성 그룹에 비해 확연히 연령이 TG 수치 기준치 초과에 영향을 줌을 알 수 있다. 흥미 있는 부분은 여성의 경우 TG와 HDL-C 반응변수 간에 상관성을 의미하는 ρ 추정치가 0에 근접함을 알 수 있다. 남성의 경우에는 TG의 증가와 HDL-C의 감소가 양의 상관관계에 있지만, 여성의 경우에는 큰 상관이 없다고 해석할 수 있다.

4. 결론

본 연구는 반응변수 간 상관성을 고려할 수 있는 결합 다단계 일반화 선형모형을 이용하여, 다변량 경시

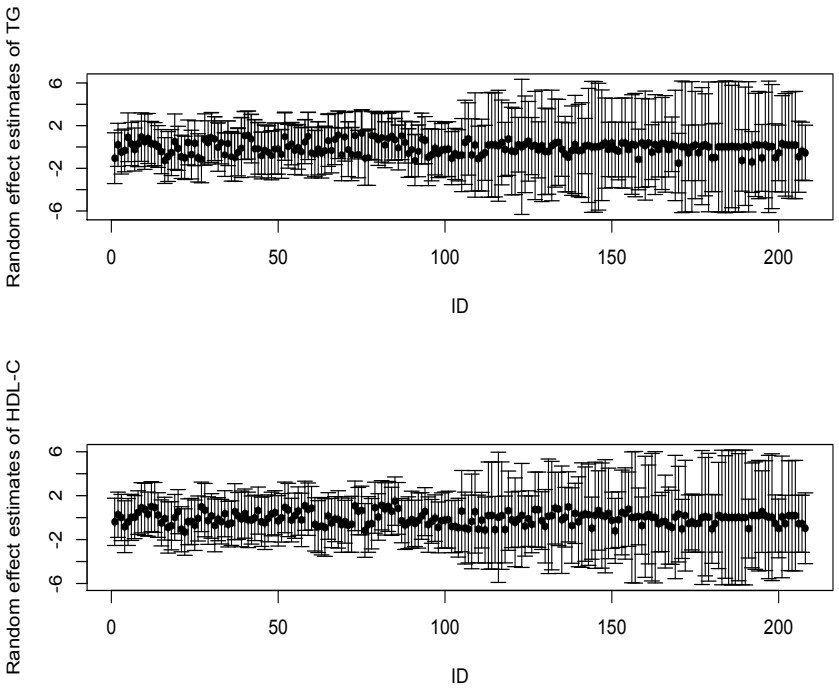

Figure 3.1. Random effects estimates and their 95% Wald intervals in Model 2

Table 3.2. Estimates of fixed parameters and variance components for male and females

	sex: male	sex: female
Response variable: TG		
intercept	−10.643 (8.303)	−23.773* (6.206)
age	2.407 (2.076)	5.749* (1.559)
drink1	−0.187 (1.088)	−0.402 (0.971)
drink2	1.357 (0.808)	−0.393 (0.551)
smoke1	0.579 (0.898)	2.589 (2.250)
smoke2	0.405 (0.831)	1.526 (1.136)
Response variable: HDL-C		
intercept	4.770 (12.569)	−4.547 (5.222)
age	−0.262 (3.162)	1.605 (1.321)
drink1	0.289 (1.662)	0.496 (0.956)
drink2	−2.469* (1.133)	−0.447 (0.454)
smoke1	−2.552 (1.349)	27.227 (1136.106)
smoke2	0.863 (1.277)	1.773 (1.234)
Variance components:		
$\log(\sigma_1^2)$	1.501* (0.353)	1.473* (0.329)
$\log(\sigma_2^2)$	2.522* (0.393)	0.551 (0.292)
ρ	0.425* (0.048)	0.001 (0.055)

Standard errors are given in parentheses.

$*$: indicates significance with 95% Wald confidence interval.

적 자료 분석을 수행하여 보았다. 한국 유전체 역학 연구에서 실시한 코호트 자료를 이용하여, 모형을 적합하고 두 반응변수에 대응되는 변량효과 간에 높은 상관성이 있음을 확인하였다. 이는 조건부 아카이케 정보 기준(cAIC)을 통해서도 변량효과 간 독립을 가정한 모형보다는 상관성이 있는 모형을 선택하고 있다. 또한 남녀 그룹별로 부분 분석을 수행해봄으로써 남녀 간의 모형 차이도 있음을 확인하였다. 전체 코호트 자료에서도 남녀 그룹별 변량효과들의 공분산 구조를 다르게 해야 할 가능성이 있다. 현재 `mdhglm` 패키지는 결측 메카니즘까지 모형에 적하고 있지는 않아 코호트 자료 전체를 분석하는 데는 어려움이 있어 결측 또는 중도탈락 메카니즘의 모형이 반영 된다면 전체 자료의 분석에 매우 유용할 것이다. 또한, JHGLM의 각 HGLM 모형이 서로 달라도 모형 적합이 가능하므로, 이진형 자료, 연속형 자료 또는 계수 자료 등의 다변량 분석이 가능하므로, 향후 보다 유연하고 다양한 분석이 기대된다.

References

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association*, **88**, 125–134.

Donohue, M., Overholser, R., Xu, R. and Vaida, F. (2011). Conditional akaike information under generalized linear and proportional hazards mixed models, *Biometrika*, **98**, 685–700.

Kim, J., Kim, E., Yi, H., Joo, S., Shin, K., Kim, J., Kimm, K. and Shin, C. (2006). Short-term incidence rate of hypertension in Korea middle-aged adults, *Journal of Hypertension*, **24**, 2177–2182.

Lee, K., Joo, Y., Yoo, J. K. and Lee, J. (2009). Marginalized random effects models for multivariate longitudinal binary data, *Statistics in Medicine*, **28**, 1284–1300.

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., Nelder, J. A. and Pawitan, Y. (2006). *Generalised Linear Models with Random Effects*, Chapman and Hall/CRC, Boca Raton.

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.

Yun, S. and Lee, Y. (2004). Comparison of hierarchical and marginal likelihood estimators for binary outcomes, *Computational Statistics and Data Analysis*, **45**, 639–650.

결합 다단계 일반화 선형모형을 이용한 다변량 경시적 자료 분석

이동환[a,1] · 유재근[a]

[a]이화여자대학교 통계학과

(2015년 3월 24일 접수, 2015년 3월 31일 수정, 2015년 3월 31일 채택)

요 약

경시적 자료는 각 환자마다 시간에 따라 반복 측정되는 코호트 연구 등에서 많이 쓰인다. 본 연구는 반응변수 간 상관성을 고려할 수 있는 결합 다단계 일반화 선형모형을 이용하여, 다변량 경시적 자료 분석을 수행하였다. 한국 유전체 역학 연구에서 실시한 코호트 자료를 적합하고 결과를 해석한다. 조건부 아카이케 정보 기준을 이용하여 모형 선택을 하고, 변량효과들의 추정치들을 설명한다.

주요용어: 결합 다단계 일반화 선형 모형, 혼합 모형, 코호트 자료, 경시적 이진 자료.

본 연구는 한국연구재단을 통해 미래창조과학부의 뇌과학원천기술개발사업으로부터 지원받아 수행되었습니다 (2014M3C7A1062896).

[1]교신저자: (120-750) 서울특별시 서대문구 대현동 11-1, 이화여자대학교 통계학과.
E-mail: donghwan.lee@ewha.ac.kr