응용통계연구 (2008)
21(6), 923-932

다변량 다수준 이항자료에 대한 일반화선형혼합모형

임화경[1] · 송석헌[2] · 송주원[3] · 전수영[4]

[1]고려대학교 통계학과; [2]고려대학교 통계학과; [3]고려대학교 통계학과;
[4]고려대학교 경제통계 산학협력단

(2008년 5월 접수, 2008년 9월 채택)

요 약

우리는 자명하지 않은 상관 구조를 갖는 복잡한 다변량 자료에 직면하는 경우가 있다. 예를 들어 군집 구조 자료의 경우 생략된 변수들이 한 개 이상의 관측값에 동시적으로 영향을 줄 수 있기 때문에 결과들 간에 상관 구조를 모형화하는 것은 추정량의 효율성과 정확한 표준오차의 계산 등의 타당한 추론을 위해서 중요하다. 관측값들 간에 종속성을 두는 표준 방법으로는 관측값들이 관찰되지 않은 어떤 변수를 공유한다고 가정하는 것인데, 이러한 가정에 대해 본 연구에서는 다수준 모형을 고려한 상관된 임의효과 모형을 적합시켰다. 추정은 준모수적 접근방법으로 임의계수 분포에 대한 모수적 가정 없이 유한혼합 EM-알고리즘을 통하여 수행되었다.

주요용어: 일반화선형혼합모형, 다수준, 상관된 임의계수, 비모수 최대우도.

1. 서론

다수준모형(multilevel model)은 계층적 또는 군집 구조의 자료를 분석하기 위한 모형으로 다양한 분야에서 사용되어진다. 예를 들면 학교 안의 학생들에 관한 교육연구, 가족 안의 아이들에 관한 가계연구, 의사나 병원 안의 환자들에 대한 의학연구 혹은 생물학적 연구 등이 있다. 여기서 하위 수준의 단위들(학생, 아이, 환자)은 상위 수준의 단위들(학교, 가족, 의사나 병원) 속에 지분(nested)되어 있다고 말할 수 있다. 동일한 개체에 대하여 일정한 시간 간격으로 반복 측정된 경시적 자료도 하위 수준인 반복 측정치들이 상위 수준인 개체들 속에 군집이 되어 다수준 자료라 볼 수 있다. 이러한 자료들의 다수준 모델링에 대한 다양한 연구가 진행되어 왔다 (Aitkin 등, 1981; Aitkin과 Longford, 1986; Goldstein, 1995; Kreft와 de Leeuw, 1998; Snijders와 Bosker, 1999; Raudenbush와 Bryk, 2002). 동일한 군집 속의 단위들은 같은 군집 중심적(cluster-specific) 효과를 공유하게 된다. 예를 들면, 같은 교실 안의 학생들은 동일한 선생님께 배우고, 그 학부모들은 주거지역이기 때문에 또는 선택에 의해 자녀를 그 학교로 보낸다. 하지만 이러한 관련된 공변량들(covariates)은 제한된 지식과 정보의 부족으로 모두 분석에 포함시키기란 불가능하다. 결과적으로 군집되어 있는 자료들 사이에 상관관계가 생겨 관측되지 않은 이질성(unobserved heterogeneity)이 발생한다. 이러한 관측되지 않은 이질성은 선형예측식에 임의

[1]교신저자: (136-701) 서울시 성북구 안암동 5-1, 고려대학교 통계학과, 박사과정 수료.
E-mail: hklim@korea.ac.kr
[2](136-701) 서울시 성북구 안암동 5-1, 고려대학교 통계학과, 교수. E-mail: ssong@korea.ac.kr
[3](136-701) 서울시 성북구 안암동 5-1, 고려대학교 통계학과, 부교수. E-mail: jsong@korea.ac.kr
[4](339-700) 충남 연기군 조치원읍 서창리 208, 고려대학교 경제통계 산학협력단, 연구교수.
E-mail: scheon@korea.ac.kr

효과(random effects)를 삽입하여 설명될 수 있는데, 반응변수의 분포가 지수족(exponential family)에 속하면 일반화선형모형(GLM)에 임의효과를 삽입하여 일반화선형혼합모형(GLMM)으로 확장할 수 있다.

다수준 이항 자료의 경우 동일한 군집 안에 지분된 반복 측정된 결과들은 급간상관(intra-class correlation)이 존재하여 과대산포(overdispersion)가 발생하는데, 이 변동은 보통 선형예측식에 임의효과를 삽입하여 일반화선형혼합모형으로 설명될 수 있다. Aitkin (1999)은 상위 수준에 일변량 정규분포를 따르는 임의효과를 삽입하여 군집되어 있는 자료들 사이에 상관관계를 설명하려고 하였지만, 이것은 개체 안의 모든 결과들이 같은 이질성을 갖도록 하는 제약이 있기 때문에 본 연구에서는 다수준 모형을 고려한 상관된 임의계수모형(correlated random coefficient models)을 통해 결과들이 일반화된 상관구조를 갖도록 하고자 한다. 한편, 임의계수 벡터의 분포는 보통 정규분포를 가정하는데, 분포에 대한 잘못된 가정으로 인하여 모수 추정의 편향이 생길 수 있기 때문에 (Aitkin, 1999), 본 연구에서는 임의계수 분포에 대한 모수적 가정 없이 유한혼합(finite mixture)을 통해 분포를 구하고 일반화선형모형의 모수도 추정하는 준모수적 접근방법(semi-parametric approach)을 제안한다.

본 연구의 구성은 다음과 같다. 2절에서는 다변량 이항자료에 대한 일반화선형혼합모형을 소개하고 임의계수 분포를 다변량 정규분포로 가정했을 때의 모수 추정 방법들에 대해서 간단히 설명한다. 3절에서는 임의효과 분포에 대한 모수적 가정을 하지 않았을 때의 비모수적 최대우도(non-parametric maximum likelihood: NPML) 추정방법에 대해 알아보고, 4절에서는 제안된 준모수적 방법의 효과를 모의실험을 통하여 평가하도록 한다. 5절에서는 제안된 모형을 실제 자료에 적용하여 분석하고, 6절에서는 본 논문의 결론을 맺는다.

2. 다변량 이항자료에 관한 일반화선형혼합모형

2.1. 공유 임의절편 모형(shared random intercept models)

i번째 개체 안에 지분된 j번째 시점의 반응이 y_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, J$인 n개 그룹화된 이항자료가 있다고 가정하자. 여기서 i번째 개체에 대한 관측값의 벡터를 $y_i = (y_{i1}, \ldots, y_{iJ})'$라 하면 반응변수 y_{ij}는 다음과 같은 이항분포를 따른다 (Aitkin, 1999).

$$y_{ij}|\,\pi_{ij} \sim \mathrm{Bin}(n_i,\, \pi_{ij}), \quad i = 1, \ldots, n,\ j = 1, \ldots, J. \tag{2.1}$$

이항 반응변수 y_{ij}에 대해 정준연결함수를 사용한 일반화선형혼합모형은 다음과 같다.

$$\mathrm{logit}(\pi_{ij}) = X_{ij}{}'\beta_j + Z_{ij}{}'b_i, \quad i = 1, \ldots, n,\ j = 1, \ldots, J, \tag{2.2}$$

여기서 $X_{ij} = (1, x_{ij1}, \ldots, x_{ijp})'$는 고정효과에 대한 $(p+1)$차원 계획행렬이고, $\beta_j = (\beta_{j0}, \ldots, \beta_{jp})'$는 고정효과에 대한 $(p+1)$차원 모수벡터이다. $Z_{ij} = (z_{ij1}, \ldots, z_{ijq})'$는 임의효과에 대한 q차원 계획행렬이고, b_i는 q차원 벡터로서 i번째 개체 안에 지분된 결과들이 공유하는 개체 중심적 임의계수를 나타내어 결과들 사이에 상관관계는 개체의 여분 이항 변동(extra-binomial variation)을 설명한다. 만약 $Z_{ij} \equiv 1$인 임의절편 모형을 고려한다면 식 (2.2)는 다음과 같이 간단히 표현될 수 있고, 여기서 b_i의 분포는 $b_i \sim N(0, \sigma^2)$을 가정한다.

$$\mathrm{logit}(\pi_{ij}) = X_{ij}'\beta_j + b_i. \tag{2.3}$$

q차원 임의벡터 b_i가 주어졌을 때의 반응변수 y_{ij}는 조건부 독립인 이항분포를 따르므로 b_i들을 장애모

수(nuisance parameter)로 다루어 그것에 대해 적분하면 다음과 같은 주변우도함수를 구할 수 있다.

$$L(\bullet) = \prod_{i=1}^{n} \int \left\{ \prod_{j=1}^{J} f(y_{ij} | \, x_{ij}, \, b_i) \right\} dG(b_i). \tag{2.4}$$

하지만 식 (2.4)의 적분은 폐쇄형(closed form)을 갖지 않으므로 가우스-에르미트 구적(Gauss-Hermite quadrature)과 같은 수치적분을 이용하여 주변우도함수를 근사시킬 수 있다.

2.2. 상관된 임의계수 모형(correlated random coefficient models)

앞의 모형은 개체 안의 모든 결과들이 같은 이질성을 갖도록 하는 제약이 있기 때문에 상관된 임의계수를 허용하여 다음과 같이 일반화 시킬 수 있다. $B_i = (b_{i1}, \ldots, b_{iJ})'$를 i번째 개체에 대한 결과들의 임의계수 벡터라 하자. b_{ij}가 주어졌을 때의 관측값 y_{ij}는 조건부 독립인 이항분포를 따르고, b_{ij}는 개체 중심적 이질성과 결과들 사이의 종속성 모두를 설명한다. 선형예측식은 다음과 같다.

$$\text{logit}(\pi_{ij}) = X'_{ij}\beta_j + Z'_{ij}b_{ij}, \quad i = 1, \ldots, n, \; j = 1, \ldots, J. \tag{2.5}$$

주변우도함수는 다음과 같이 임의계수 벡터에 대해 적분함으로서 구할 수 있다.

$$L(\bullet) = \prod_{i=1}^{n} \int \prod_{j=1}^{J} f(y_{ij} | \, x_{ij}, \, b_{ij}) \, dG(B_i). \tag{2.6}$$

앞에서와 마찬가지로 위의 적분은 폐쇄형을 갖지 않으므로 수치적분을 통해서 근사시켜야 한다. 만약 $q = 1$에 대하여 임의계수 B_i의 분포가 $B_i \sim \text{MVN}(0, \sum)$을 따른다고 가정한다면 위 식의 근사는 다음과 같은 재모수화(reparametrization)를 통해서 더 쉽게 행할 수 있다 :

$$B_i = \sum^{\frac{1}{2}} B_i^*, \tag{2.7}$$

여기서 $\sum^{1/2}$은 하삼각행렬(lower triangular matrix)로서 촐레스키 분해(Cholesky decomposition) $\sum = \sum^{1/2} \sum^{1/2'}$에 의하여 구할 수 있고, $B_i^* \sim \text{MVN}(0, I)$ 을 따른다. 일반성에 손실없이 ' * '를 제거하여 $B_i^* = B_i$라 놓으면, 식 (2.5)의 선형예측식은 다음과 같이 수정될 수 있다.

$$\text{logit}(\pi_{ij}) = X'_{ij}\beta_j + Z'_{ij}\sum_i^{\frac{1}{2}} B_i = \left[\, X'_{ij}, \; (Z_{ij} \otimes B_i)' \, \right] \begin{bmatrix} \beta_j \\ \sigma \end{bmatrix}, \tag{2.8}$$

$$\text{여기서} \quad \sigma = \text{vec}\left(\sum^{\frac{1}{2}} \right).$$

가우스-에르미트 공식을 이용하여 근사시킨 주변우도함수는 다음과 같다.

$$L(\bullet) = \prod_{i=1}^{n} \sum_{m=1}^{M} w_m \left\{ \prod_{j=1}^{J} f(y_{ij} | \, x_{ij}, \, B_m) \right\}, \tag{2.9}$$

여기서 $B_m = (b_{m1}, \ldots, b_{mJ})'$는 구적 위치벡터(points)이고 관련된 질량(mass)은 $w_m = \prod_{j=1}^{J} w_{mj}$이다. b_m와 w_m의 값은 $m \leq 20$의 경우 Abramowitz와 Stegun (1972), $m > 20$의 경우는 Golub와 Welsch (1969)에 있다.

우도함수를 근사시킨 후엔 Quasi-Newton과 같은 방법으로 우도함수를 최대화시켜 모수를 추정하거나 Aitkin (1999)의 방법에 따라 EM-알고리즘을 이용하여 유한혼합모형을 통해 모수를 추정할 수도 있는데 방법은 아래와 같다. $\delta' = (\beta_1', \ldots, \beta_J', \sigma)$를 완전모수벡터(complete parameter vector)로 두고 만약 $(\partial \log f_{im})/\partial\delta = \sum_{j=1}^{J} \partial \log f_{ijm}/\partial\delta$, 여기서 $f_{im} = \prod_{j=1}^{J} f(y_{ij}|\, x_{ij}, b_{mj}) = \prod_{j=1}^{J} f_{ijm}$라 하면 조건부 로그우도함수의 δ에 대한 스코어 함수는 다음과 같다.

$$\frac{\partial \log L(\delta)}{\partial\delta} = \frac{\partial \ell(\delta)}{\partial\delta} = \sum_{i=1}^{n}\sum_{m=1}^{M}\left(\frac{w_m f_{im}}{\sum_{m=1}^{M} w_m f_{im}}\right)\frac{\partial \log f_{im}}{\partial\delta} = \sum_{i=1}^{n}\sum_{m=1}^{M} z_{im}\frac{\partial \log f_{im}}{\partial\delta}, \tag{2.10}$$

여기서 z_{im}는 성분 m로부터 나온 i번째 개체에 대한 사후확률(posterior probability)을 나타낸다. 스코어함수가 0과 같다고 둔 우도방정식은 가중치를 가진 보통의 일반화선형모형의 가중합 형태이고 주어진 가중치들에 대해 이 방정식을 풀고 현재의 모수 추정치로부터 구한 가중치들을 갱신하여 다시 방정식을 푸는 EM-알고리즘을 수행하여 모수 추정을 할 수 있다. SAS의 PROC NLMIXED는 적응 가우시안 구적(adaptive Gaussian quadrature) 또는 1차 테일러 시리즈 근사와 같은 방법을 통해 적분을 근사시키고, Quasi-Newton이나 Newton-Raphson과 같은 여러 가지 최적화 알고리즘 중에 하나를 선택하여 EM-알고리즘 없이 모형을 적합시킬 수 있다. 하지만 임의효과가 포함된 복잡한 모형인 경우에는 사용될 수 없다.

3. 비모수적 최대우도 추정방법(NPML)

다양한 공분산구조를 포함할 수 있는 임의효과에 대해 일반적으로 다변량 정규성을 가정하여 분석한다. 이것은 임의효과의 분포를 명시할 수 없기 때문인데 본 연구는 준모수적 접근을 통해 임의효과의 분포를 추정하고자 한다. 준모수적 접근이라 함은 기본 모형이 모수적 접근과 같은 다항식의 형태를 가지면서 임의효과 분포의 모수들(즉, 위치벡터와 질량)의 추정을 위해 비모수적 접근방법을 포함하기 때문이다. 표준 가정은 다음과 같다.

- K는 고정된 상수이고 미지이다.
- 반응변수 $y_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, J$는 조건부 독립인 이항분포를 따른다.

$$\begin{aligned} y_{ij}|\,\pi_{ij} &\sim \mathrm{Bin}(n_i, \pi_{ij}), \\ &\text{여기서 } \mathrm{logit}(\pi_{ij}) = X_{ij}'\beta_j + Z_{ij}'b_{ij}. \end{aligned} \tag{3.1}$$

- 관측되지 않은 다변량 임의계수 벡터 $B_i = (b_{i1}, \ldots, b_{iJ})$에 대한 분포 $G(B_i)$는 위치벡터 $B_k = (b_{k1}, \ldots, b_{kJ}),\ k = 1, \ldots, K$와 질량 $P(B_i = B_k) = p_k,\ \sum_{k=1}^{K} p_k = 1$을 갖는 이산 확률질량함수이다.

$(y_i, X_i, B_i) \equiv (y_{i1}, \ldots, y_{iJ},\ x_{i1}, \ldots, x_{iJ},\ b_{i1}, \ldots, b_{iJ})',\ i = 1, \ldots, n,\ j = 1, \ldots, J$는 혼합모형의 완전자료(complete data)이고 여기서 (y_i, X_i)은 관측된 자료, B_i는 관측되지 않은 자료이다. 관측되지 않은 성분 지시변수 벡터 $z_i = (z_{i1}, \ldots, z_{iK})',\ i = 1, \ldots, n,\ k = 1, \ldots, K$를 다음과 같이 놓자.

$$z_{ik} = \begin{cases} 1, & \text{if } B_i = B_k, \\ 0, & \text{otherwise.} \end{cases}$$

i번째 개체가 혼합분포의 k번째 성분에 속할 사전확률(prior probability)을 $P(z_{ik}=1)=p_k$라 하면, 완전자료에 대한 로그우도함수는 다음과 같다.

$$\ell_{com}(\bullet)=\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log(p_k)+\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log(f_{ik}), \tag{3.2}$$

$$\text{여기서 } f_{ik}=\prod_{j=1}^{J} f(y_{ij}|\,x_{ij},\,b_{kj}).$$

E-단계에서는 완전자료 로그우도함수의 z_{ik}가 관측되지 않았으므로 이것을 $y_i=(y_{i1},\ldots,y_{iJ})'$와 현재의 모수 추정치 $\delta^{(r)}=(\beta_1,\ldots,\beta_J,\sigma,B_1,\ldots,B_K)^{(r)}$가 주어졌을 때의 z_{ik}에 대한 조건부 기대값인 w_{ik}로 대체함으로서 관측된 자료에 대한 로그우도함수를 구할 수 있다.

$$\widehat{z_{ik}}\left(\delta^{(r)}\right)=w_{ik}{}^{(r)}=\frac{p_k{}^{(r)}\prod_{j=1}^{J} f\left(y_{ij}|\,x_{ij},\,B_k{}^{(r)}\right)}{\sum_{k=1}^{K} p_k{}^{(r)}\prod_{j=1}^{J} f\left(y_{ij}|\,x_{ij},\,B_k{}^{(r)}\right)}, \tag{3.3}$$

여기서 $\widehat{z_{ik}}(\delta^{(r)})=w_{ik}^{(r)}$는 i번째 개체가 혼합분포의 k번째 성분에 속할 사후확률을 가리킨다. 관측값 y_i가 주어졌을 때 완전자료 로그우도함수의 조건부 기대값은 다음과 같다.

$$Q\left(\delta|\,\delta^{(r)}\right)=E_{\delta^{(r)}}(\ell_{com}(\bullet)|\,y_i)=\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\log(p_k)+\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\log(f_{ik}). \tag{3.4}$$

M-단계는 사후확률 $\widehat{z_{ik}}=\widehat{z_{ik}}(\delta^{(r)})$이 주어졌을 때 모수에 대한 새로운 최대우도추정치 $\delta^{(r+1)}$을 구하는 단계로서 $Q(\bullet)$을 δ에 대하여 최대화 시킨다.

$$\frac{\partial Q}{\partial p_k}=\sum_{i=1}^{n}\frac{\widehat{z_{ik}}}{\widehat{p_k}}-\frac{\widehat{z_{iK}}}{\widehat{p_K}}=0, \tag{3.5}$$

$$\frac{\partial Q}{\partial \delta}=\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}{}^{(r)}\frac{\partial\log(f_{ik})}{\partial\delta}=0. \tag{3.6}$$

식 (3.5)를 최대화 하면 $\widehat{p_k}{}^{(r)}=1/n\sum_{i=1}^{n} w_{ik}{}^{(r)}$을 얻을 수 있고, 식 (3.6)은 폐쇄형 해를 구할 수 없으므로 Quasi-Newton과 같은 수치적 방법을 이용하여 구할 수 있다. 위의 E-단계와 M-단계를 반복 수행하면서 관측된 로그우도함수 ℓ이 다음을 만족시키면 실행을 중지한다. 본 연구에서는 최적정지규칙으로 $|\ell^{(r+1)}-\ell^{(r)}|/|\ell^{(r)}|<10^{-8}$을 사용하였다. 혼합성분의 수 K를 결정하기 위해서는 일단 고정된 K에 대해서 계산을 한 후, $K+1$로 성분 수를 증가시켜 나가면서 수행할 수 있다. 모형 비교는 AIC, BIC와 같은 벌점우도(penalized likelihood) 기준을 사용하거나 $H_0: K=K_1$ $vs.$ $H_1: K=K_2$, $K_1<K_2$에 대한 $-2\{\ell(\widehat{\delta}_K)-\ell(\widehat{\delta}_{K+1})\}$을 계산하여 우도비 검정(likelihood ratio test)을 할 수도 있다.

4. 모의실험

제안된 방법의 성능을 평가하기 위해서 $J=2$인 경우의 이항자료를 가지고 모의실험을 실시하였다. 우선 결과들은 조건부 독립성을 만족한다고 가정하고 다음과 같이 이항분포로부터 $n_i=30$인 표본들을 생

표 4.1. $(\beta_{10}, \beta_{11}, \beta_{20}, \beta_{21}) = (1, 0.5, 0.5, 1)$, $K = 2$인 경우 모의실험 결과

n	ρ	$\hat{\beta}_{10}$ (SE)	$\hat{\beta}_{11}$ (SE)	$\hat{\beta}_{20}$ (SE)	$\hat{\beta}_{21}$ (SE)	$\hat{\rho}_{(u_1,u_2)}$ (SE)	Log-Likelihood $\bar{\ell}$
50	−0.4	0.9461 (0.4181)	0.5412 (0.5228)	0.4853 (0.4696)	1.2112 (0.4867)	−0.5163 (0.4633)	−301.62
	0.3	0.9614 (0.4980)	0.5394 (0.4627)	0.4811 (0.5096)	1.3275 (0.5065)	0.3148 (0.3623)	−303.60
	0.6	0.9738 (0.4311)	0.4705 (0.4916)	0.5319 (0.4523)	1.1812 (0.4827)	0.7144 (0.4284)	−297.03
100	−0.4	0.9782 (0.3352)	0.4874 (0.3494)	0.5148 (0.3749)	1.1119 (0.4118)	−0.4889 (0.3052)	−608.08
	0.3	0.9744 (0.4087)	0.5182 (0.4217)	0.4924 (0.3958)	1.2213 (0.4349)	0.2887 (0.3506)	−611.75
	0.6	0.9792 (0.3535)	0.5223 (0.3778)	0.4826 (0.3973)	1.1134 (0.3672)	0.6578 (0.3428)	−599.22
200	−0.4	0.9822 (0.3309)	0.4979 (0.2821)	0.5067 (0.3194)	0.9841 (0.2878)	−0.4692 (0.2224)	−1207.40
	0.3	0.9865 (0.2573)	0.4951 (0.3225)	0.4953 (0.2896)	1.0819 (0.3062)	0.2879 (0.2680)	−1215.63
	0.6	0.9846 (0.3232)	0.5163 (0.2432)	0.5112 (0.2534)	0.9882 (0.2860)	0.6451 (0.2827)	−1197.25
500	−0.4	1.0173 (0.2400)	0.5137 (0.2314)	0.4969 (0.2023)	1.0127 (0.2276)	−0.4513 (0.2873)	−3031.94
	0.3	0.9965 (0.2613)	0.5008 (0.2314)	0.5031 (0.2656)	0.9970 (0.2068)	0.3028 (0.1982)	−3056.27
	0.6	0.9985 (0.2377)	0.5043 (0.2128)	0.4976 (0.2039)	1.0155 (0.2273)	0.6223 (0.1975)	−2995.77

성하였다.

$$
\begin{aligned}
y_{i1}|\,\pi_{i1} &\sim \mathrm{Bin}(n_i,\, \pi_{i1}),\\
y_{i2}|\,\pi_{i2} &\sim \mathrm{Bin}(n_i,\, \pi_{i2}),\\
\text{여기서 } \mathrm{logit}(\pi_{i1}) &= \beta_{10} + \beta_{11}x_i + b_{i1},\\
\mathrm{logit}(\pi_{i2}) &= \beta_{20} + \beta_{21}x_i + b_{i2}.
\end{aligned}
\tag{4.1}
$$

개체의 수는 $n = 50, 100, 200, 500$, 공변량 x_i는 $N(0, 1)$에서 추출하였고, 잠재변수는 $\sigma_1 = \sigma_2 = 1$, $\sigma_{12} = \rho = (-0.4, 0.3, 0.6)$의 각각의 경우에 대해 이변량 정규분포로부터 다음과 같이 자료를 생성하였다.

$$
\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \sim \mathrm{BVN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix},\ \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right). \tag{4.2}
$$

모수 벡터에 대한 참값은 $(\beta_{10}, \beta_{11}, \beta_{20}, \beta_{21}) = (1, 0.5, 0.5, 1)$으로 두었고, 혼합성분의 수 $K = 2, 3, 4$에 대해서 2.2절의 상관된 임의계수 모형을 적합하였다. 혼합분포의 성분의 수와 일치하는 가우시안 구적의 위치벡터와 질량을 초기값으로 사용하였고, 최상의 로그우도를 모형 선택의 기준으로 삼았다. 표준오차 추정치는 모수적 붓스트랩을 통해 구할 수 있었다. 즉, 붓스트랩 표본으로 모형을 적합시

표 5.1. 방사선 실험 세포 자료

Occasion(i)	Dish(j)	Surviving cells(y_{ij})	Occasion(i)	Dish(j)	Surviving cells(y_{ij})
1	1	178	6	1	115
1	2	193	6	2	130
1	3	217	6	3	133
2	1	109	7	1	200
2	2	112	7	2	189
2	3	115	7	3	173
3	1	66	8	1	88
3	2	75	8	2	76
3	3	80	8	3	90
4	1	118	9	1	121
4	2	125	9	2	124
4	3	137	9	3	136
5	1	123			
5	2	146			
5	3	170			

켜 얻은 표본 공분산 행렬에 의해 모수 추정치의 공분산 행렬을 근사시켜 얻을 수 있었다.

$$\mathrm{cov}\left(\hat{\delta}\right) = \sum_{b=1}^{B} \frac{\left(\hat{\delta} - \bar{\hat{\delta}}\right)\left(\hat{\delta} - \bar{\hat{\delta}}\right)'}{B-1}, \quad \text{여기서 } \bar{\hat{\delta}} = \sum_{b=1}^{B} \frac{\hat{\delta_b}}{B}. \tag{4.3}$$

표 4.1은 표본의 수 $n = 50, 100, 200, 500$에 대하여 모수가 $(\beta_{10}, \beta_{11}, \beta_{20}, \beta_{21}) = (1, 0.5, 0.5, 1)$, $\rho = (-0.4, 0.3, 0.6)$일 때, 모의실험을 500번 반복하여 모수를 추정한 결과이다. $K = 2, 3, 4$를 고려하였는데, 최상의 로그우도를 기준으로 미지의 혼합분포 $G(\bullet)$를 추정하기 위하여 사용된 성분의 수는 $K = 2$가 적당하여 이 결과만 제공하였다. $\rho = -0.4$일 때 $\hat{\beta}_{10}$의 값은 표본의 수가 증가함에 따라 참값인 1에 근접하게 추정되었다. 이는 $\rho = 0.3$이나 0.6인 경우에도 동일하게 나타나 표본의 수가 증가할수록 상관계수 ρ의 값에 관계없이 모수에 대한 명확하고 일치성 있는 추정 결과를 얻을 수 있었고, 이와 같은 경향은 $\hat{\beta}_{11}$, $\hat{\beta}_{20}$ 그리고 $\hat{\beta}_{21}$에서도 비슷하게 나타났다. 표본의 수가 증가하면 모수들의 표준오차(SE)는 점점 줄어드는 경향을 보였다. 상관계수가 $\rho = 0.3$일 때의 $\hat{\rho}$이 가장 참값에 가깝게 추정되었다.

5. 사례연구

5.1. 자료 및 모형 설명

표 5.1은 방사선에 노출된 암세포들의 생존 여부를 측정한 실험결과이다 (Schall, 1991). 한 접시에 400개씩 세포를 담고 3개의 접시를 동시에 방사선에 노출시킨 다음 여전히 살아있는 세포들의 개수를 세었고, 이 과정을 9번 반복하여 총 27개의 접시에 대해서 암세포들의 생존 여부를 나타내는 이항 관측의 결과를 얻었다. 이 자료의 관심은 여분 이항 변동의 존재 여부였는데, 27개의 자료를 독립인 이항 관측으로 다루면 자유도가 26인 Pearson χ^2 통계량은 470.34이고, 분산 추정치는 $\hat{\sigma}^2 = 18.09$으로서 자유도인 1보다 상당히 커 과대산포를 의심할 수 있다. 이 과대산포를 설명하기 위한 한 가지 방법은 시점들(occasion) 또는 접시들(dish) 사이에 암세포들의 생존에 관한 상관성, 즉 이질성이 있다고 가정하는 것인데, 이러한 관측되지 않은 이질성은 선형예측식에 임의효과를 삽입하여 설명될 수 있다. Schall (1991)은 개체들 사이에 변동이 있고 접시들 사이에도 변동이 있음을 보였지만, 개체와 접시의 임의효과

표 5.2. 혼합 이항분포를 통해 구한 추정치들과 동질적 모집단 추정치의 비교

# of Mixing Components		Location parameter $\hat{\beta}_{kj}(\hat{\pi}_{kj})$ $J=1$	$J=2$	$J=3$	Weighting Prob.	Rho $\hat{\rho}$	Log-Likelihood $\bar{\ell}$
동질적(Homogeneous) 모집단							
$K=1$	Clust1	−0.84(0.30)	−0.77(0.32)	−0.66(0.34)	1.00	0.997	−120.50
이질적(Heterogeneous) 모집단							
$K=2$	Clust1	−0.61(0.35)	−0.64(0.35)	−0.56(0.36)	0.43	0.992	−120.22
	Clust2	−0.35(0.41)	−0.20(0.45)	−0.07(0.48)	0.57		
$K=3$	Clust1	−0.63(0.35)	−0.66(0.34)	−0.59(0.36)	0.37	0.997	−118.92
	Clust2	−0.41(0.40)	−0.32(0.42)	−0.23(0.44)	0.38		
	Clust3	−0.26(0.43)	−0.05(0.49)	0.13(0.53)	0.25		

가 독립임을 가정하여 개체와 접시의 효과를 동시에 분석에 포함시키지는 못했다. 따라서 본 연구는 패널구조의 형태로 개체내의 결과들 간에 상관성을 고려하였으며, 또한 자료에 알맞은 임의효과의 분포도 추정을 하였다. 분석 모형은 다음과 같다.

$$\begin{aligned} y_{i1}|\,\pi_{i1} &\sim \text{Bin}(400,\,\pi_{i1}), & & \text{logit}\,(\pi_{i1}) = \beta_{i1} + b_{i1}, \\ y_{i2}|\,\pi_{i2} &\sim \text{Bin}(400,\,\pi_{i2}), & \text{여기서} \quad & \text{logit}\,(\pi_{i2}) = \beta_{i2} + b_{i2}, \\ y_{i3}|\,\pi_{i3} &\sim \text{Bin}(400,\,\pi_{i3}), & & \text{logit}\,(\pi_{i3}) = \beta_{i3} + b_{i3}, \end{aligned}$$

$$\begin{pmatrix} b_{i1} \\ b_{i2} \\ b_{i3} \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix} \right). \tag{5.1}$$

5.2. 분석 결과

EM-알고리즘을 통해서 구한 추정 결과는 표 5.2에 있다. 혼합분포의 개수 K에 따른 위치모수 β_{kj}(또는 π_{kj}), 가중치 p_k 그리고 상관계수 ρ와 로그우도 값을 제시하였다. 이질적 모집단을 가정한 경우 가장 적당한 부분 모집단의 개수는 $K=3$으로 추정되었고, $K=1$인 동질적 모집단의 경우와 비교해봤을 때 로그우도 값이 약 1.58정도 차이가 나는 것을 볼 수 있어 이질성이 있다고 가정한 경우가 모집단을 더 잘 설명해준다는 것을 알 수 있었다. $K=1$인 동질적 모집단에 비해 혼합분포의 수가 $K=2,3$으로 증가할수록 위치모수는 큰 값을 가졌으며, 각 혼합분포의 가중치들은 거의 균등하였다. 방사선 실험 세포 자료의 개수는 적은데 혼합분포의 수 K가 증가할수록 추정해야 할 모수의 개수는 많아지므로, 본 분석에서는 혼합분포의 개수로 $K=4$ 이상은 고려하지 않았다.

6. 결론

같은 개체 안에 반복 측정된 결과들이 있는 다변량 자료의 경우 결과들은 보통 상관되어 있다. 이 상관구조를 모형화 하는 것은 추정량의 효율성과 정확한 표준오차의 계산 등의 타당한 추론을 위해서 중요하다. 본 연구는 다수준 모형을 고려한 상관된 임의계수모형을 통해 결과들 간의 상관성을 정의하였고, 추정은 임의계수 분포에 대한 모수적 가정 없이 유한혼합 EM알고리즘을 통하여 수행하였다. 결과들 간의 상관성 및 회귀계수들의 정확한 추정치를 구할 수 있었으며, 임의계수의 분포 역시 추정할 수 있었다.

준모수적 접근은 임의계수 분포에 대한 잘못된 가정으로 인하여 모수추정의 편향이 생기는 경우에 모수적 접근방법보다 우수할 것이다.

베이지안 방법으로도 모수 추정이 가능한데, standard reversible-jump (Green, 1995) 또는 trans-dimensional (Petris와 Tardella, 2003) MCMC 알고리즘을 사용할 수도 있다. 추후 본 논문에서 제안한 모형을 다른 추정방법을 사용하여 분석하는 것도 의미가 있을 것이다.

최대우도 추정치의 공분산 행렬에 대한 추정치는 여전히 매우 어려운 과제로 남는다. 사실 관측된 정보행렬에 기초한 표준오차는 너무나 불안정하다 (McLachlan과 Peel, 2000). 이 문제는 붓스트랩 표본으로 모형을 적합시켜 얻은 표본 공분산 행렬에 의해 모수 추정치의 공분산 행렬을 근사시켜 해결될 수는 있었으나 상당히 많은 계산량이 요구된다.

감사의 말씀

논문을 지도해 주시고 연구자의 자세를 가르쳐 주셨던 故 송석헌 교수님께 이 논문을 바칩니다.

참고문헌

Abramowitz, M. and Stegun, I. A. (1972). *Handbook of Mathematical Functions*, Dover, New York.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models, *Biometrics*, **55**, 117–128.

Aitkin, M., Anderson, D. and Hinde, J. (1981). Statistical modelling of data on teaching styles, *Journal of the Royal Statistical Society, Series A*, **144**, 419–461.

Aitkin, M. and Longford, N. T. (1986). Statistical modelling issues in school effectiveness studies, *Journal of the Royal Statistical Society, Series A*, **149**, 1–43.

Goldstein, H. (1995). *Multilevel Statistical Models*, Edward Arnold, London.

Golub, G. H. and Welsch, J. H. (1969). Calculation of Gaussian quadrature rules, *Mathematics of Computation*, **23**, 221–230.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika*, **82**, 711–732.

Kreft, I. G. G. and de Leeuw, J. (1998). *Introducing Multilevel Modeling*, Sage Publications, London.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*, John Wiley & Sons, New York.

Petris, G. and Tardella, L. (2003). A geometric approach to transdimensional MCMC, *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, **31**, 469–482.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd Edition, Sage, London.

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

Snijders, T. A. B. and Bosker, R. J. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modelling*, Sage, London.

Generalized Linear Mixed Model for Multivariate Multilevel Binomial Data

Hwa-Kyung Lim[1] · Seuck-Heun Song[2] · Juwon Song[3] · Sooyoung Cheon[4]

[1]Dept. of Statistics, Korea University;
[2]Dept. of Statistics, Korea University;
[3]Dept. of Statistics, Korea University;
[4]KU Industry-Academy Cooperation Group Team of Economics and Statistics, Korea University;

(Received May 2008; accepted September 2008)

Abstract

We are likely to face complex multivariate data which can be characterized by having a non-trivial correlation structure. For instance, omitted covariates may simultaneously affect more than one count in clustered data; hence, the modeling of the correlation structure is important for the efficiency of the estimator and the computation of correct standard errors, *i.e.*, valid inference. A standard way to insert dependence among counts is to assume that they share some common unobservable variables. For this assumption, we fitted correlated random effect models considering multilevel model. Estimation was carried out by adopting the semiparametric approach through a finite mixture EM algorithm without parametric assumptions upon the random coefficients distribution.

Keywords: GLMM, multi-level, correlated random effects, NPML.

[1]Corresponding Author: Doctoral student, Dept. of Statistics, Korea University, Anam-Dong 5, Sungbuk-Gu, Seoul 136-701, Korea. E-mail: hklim@korea.ac.kr
[2]Professor, Dept. of Statistics, Korea University, Anam-Dong 5, Sungbuk-Gu, Seoul 136-701, Korea. E-mail: ssong@korea.ac.kr
[3]Associate Professor, Dept. of Statistics, Korea University, Anam-Dong 5, Sungbuk-Gu, Seoul 136-701, Korea. E-mail: jsong@korea.ac.kr
[4]Research Professor, KU Industry-Academy Cooperation Group Team of Economics and Statistics, Korea University, Jochiwon 339-700, Korea. E-mail: scheon@korea.ac.kr