Journal of the Korean Data & Information Science Society
2014, **25**(6), 1263–1272

http://dx.doi.org/10.7465/jkdi.2014.25.6.1263
한국데이터정보과학회지

주변화 변량효과모형의 조사 및 고찰[†]

전주영[1] · 이근백[2]

[1]국립암센터 · [2]성균관대학교 통계학과

접수 2014년 7월 29일, 수정 2014년 9월 1일, 게재확정 2014년 9월 24일

요 약

경시적 범주형자료 (longitudinal categorical data)는 의학, 보건학, 그리고 사회과학에서 많이 발생하는 자료이다. 이러한 자료는 반복측정으로 인한 결과치들의 상관관계를 설명하면서 공변량의 효과를 설명해야 한다. 이 논문에서 모집단에 대한 공변량의 효과를 추정하면서 우도함수에 기초한 모형인 주변화 변량효과모형 (marginalized random effects model)을 소개하고, 그 모형의 어떻게 발전했는지를 고찰한다. 그리고 실제 자료를 이용하여 제시된 모형을 설명한다.

주요용어: 경시적자료, 공변량, 반복측정, 조건부모형, 주변모형.

1. 머리말

경시적 범주형자료 (longitudinal categorical data)는 의학, 보건학, 그리고 사회과학에서 많이 발생하는 자료이다. 경시적자료는 개체 (subject)의 반복된 측정에 의한 상관관계를 가지고 있다. 따라서 단순한 일반화선형모형을 이용하여 모형을 추정하면 그 추정치는 편의 (bias)가 발생할 수 있다 (Daniels와 Hogan, 2008). 따라서 이러한 편의를 줄이기 위해 다양한 모형들이 제안되었는데 그 중 특히 많이 사용되는 모형들이 조건부모형 (conditional models)과 주변모형 (marginal models)으로 분류할 수 있다 (Daniels와 Hogan, 2008).

조건부 모형은 주로 개체특성적 효과 (subject-specific effect)에 관심이 있을 때에 주로 사용되고, 주변모형은 모집단의 평균적인 효과 (population-averaged effect)에 주된 관심이 있을 때 사용된다. 우선 조건부모형은 변량효과변수 (random effect variable)를 이용하여 반복 측정된 결과치들의 상관관계를 설명하는 일반화선형혼합모형 (generalized linear mixed models; GLMM) (Breslow와 Clayton, 1993)과 그전 시간에 관찰된 결과치를 이용하는 마코프 (Markov) 구조를 이용하는 전이모형 (transition models)이라 한다 (Diggle 등, 2002). 이 두 모형은 모두 응답변수 (response variable)의 조건부평균과 공변량 (covariate)의 관계를 간접적으로 설명하는 모형이다. 이러한 조건부모형은 최대우도 (maximum likelihood)방법으로 모수들의 추정치를 찾을 수 있다. 따라서 최대우도치 (maximum likelihood estimator)의 특성인 일치성 (consistency)과 점근적 정규분포성 (asymptotic normality)을 가지는 특성이 있다.

주변모형은 모집단에 대한 공변량효과 (covariate effect)에 관심이 있을 때 주로 사용하는 모형이다. 모수추정의 방법으로는 주변모형에서의 평균과 분산, 주변 확률밀도함수(marginal probability density

[†] 이 논문은 2012년도 정부 (미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2012R1A1A1004002).
[1] (110-745) 경기도 고양시 일산동구 일산로 323, 국립암센터 중앙암등록사업과, 연구원.
[2] 교신저자: (110-745) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과, 부교수.
E-mail: keunbaik@skku.edu

function), 그리고 가상관행렬 (working correlation matrix)만을 이용하는 GEE (generalized estimating equation)를 이용한다 (Liang와 Zeger, 1986). 하지만 GEE를 이용한 주변모형의 경우 경시적자료에서 자주 발생하는 임의결측치 (missing at random)가 있을 때는 바로 사용할 수 없다 (Daniels와 Hogan, 2008). 또한 최대우도원리를 기초로 한 방법이 아니므로 우리가 흔히 쓰는 우도비검정과 벌점에 기초한 모형선택 기준들 (penalized model selection criterion)인 AIC 또는 BIC 등을 사용할 수 없는 한계가 있다. 이런 단점들을 극복하면서 주변모형이 가지는 장점들을 살리는 모형들이 개발되었는데, 그 중에서 주변화모형 (marginalized moels)이 있다 (Heagerty, 1999, 2002). 주변화모형은 우도원리를 기초로 두 개의 부모형 (sub models)을 가진다. 공변량의 모집단의 평균적인 효과를 설명하기 위한 주변평균모형 (marginal mean model)과 반복 측정된 결과치 들의 상관관계를 설명하는 의존모형(dependence model)으로 구분된다. 의존모형은 변량효과 또는 마코프 구조를 이용한 모형으로 구분되며, 변량효과를 이용하는 것을 주변화 변량효과모형 (marginalized random effects model)이라고 하고, 마코프 구조를 이용하는 것을 주변화 전이모형 (marginalized transition model)이라고 한다. 이 두 모형은 의존모형에서 조건부모형의 특성을 그리고 주변평균모형에서 주변모형의 특성을 가지므로 조건부모형과 주변모형의 장점을 모두 가질 수 있다. 그 장점들은 아래와 같다. 첫째, 조건부모형처럼 최대우도원리를 이용할 수 있다. 따라서 최대우도치의 장점들인 일치성과 점근적정규성을 가진다. 둘째, 주변모형과 같이 모집단의 평균적인 효과를 직접적으로 설명할 수 있다. 셋째, 주변평균모형에서 독립변수의 모집단의 평균적인 효과의 해석 시에 의존모형의 형태에 영향을 받지 않으므로 잘못된 변량효과의 분포로 인한 편의의 영향이 적다 (Heagerty와 Zeger, 2000; Heagerty와 Kurland, 2001; Heagerty, 2002; Lee와 Daniels, 2007; Lee와 Mercante, 2010; Lee 등, 2011). 마지막으로 무시할 수 있는 결측치 (ignoreable missingness)가 있을 경우, 주변화모형은 우도원리에 기초한 모형이므로 모수추정 시에 그 영향을 무시할 수 있다. 이러한 장점 때문에 경시적 범주형자료의 분석에서 주변화모형이 이용되고 있다.

주변화 변량효과모형이 Heagerty (1999)에 의해 제안된 이후 Lee와 Daniels (2008)이 경시적 순서자료의 분석을 위한 모형으로 확장되었고, Lee 등 (2011)에서는 경시적 명목자료 (longitudinal nominal data)분석을 위해서 확장되었다. 위에 제시된 주변화 변량효과모형의 변량효과의 공분산행렬은 모두 단순한 $AR(1)$구조를 가정하고 있다. 하지만 이 가정은 강한 가정일 때도 있다. 따라서 변량효과의 공분산행렬의 구조를 일반적인 $AR(p)$로 가정하면서 이분산성을 만족하는 모형이 수정 콜레스키분해 (modified Cholesky decompositon; MCD) 방법이 개발되었다 (Pourahmadi, 1999). 이 논문에서 위에 제시된 논문들에서 제안된 모형들 살펴보고자 한다.

이 논문의 구성은 아래와 같다. 2절에서는 경시적 이진자료를 위한 주변화 변량효과모형에 대한 연구들을 살펴본다. 그리고 3절에서는 경시적 다범주형 자료분석을 위한 주변화 변량효과모형에 대한 연구들을 살펴본다. 그리고 4절에서는 복지 패널자료를 이용하여 앞장에서 제시된 모형에 적용시켜 본다. 마지막으로 5절에서 요약과 결론을 제시한다.

2. 경시적 이진자료 분석

여기에서 우선 경시적 이진자료 (longitudinal binary data) 분석을 위한 주변화모형인 주변화 변량효과모형을 설명하고, 그리고 변량효과공분산행렬의 일반화에 대해 설명한다. 우선 주변화모형이 개발된 배경은 다음과 같다.

주변모형에서 모수추정의 방법으로는 주변모형에서의 평균과 분산, 주변 확률밀도함수 (marginal probability density funcion), 그리고 가상관행렬 (working correlation matirx)만을 이용하는 GEE를 이용한다(Liang 과 Zeger, 1986). 하지만 GEE를 이용한 주변모형의 경우 경시적자료에서 자주 발생하

는 임의결측치가 있을 때는 바로 사용할 수 없다 (Daniels와 Hogan, 2008). 또한 최대우도원리를 기초로 한 방법이 아니므로 우리가 흔히 쓰는 우도비검정과 벌점에 기초한 모형선택 기준(penalized model selection criterion)인 AIC 또는 BIC 등을 사용할 수 없는 한계가 있다. 이런 단점을 보완하고 주변모형이 가지는 장점들을 살려서 개발된 모형이 주변화모형이다.

이제 자료의 형태를 보면 다음과 같다. $Y_i^T = (Y_{i1}, \cdots, Y_{n_i})$를 i번째 개체의 이진자료의 응답변수벡터로 한다. 그리고 x_{it}는 i번째 개체의 t번째 독립변수로 $p \times 1$벡터로 한다.

2.1. 주변화 변량효과모형

주변화 변량효과모형은 Heagerty (1999)에 의해서 제안되었고, 그 모형의 형태는 아래와 같다.

$$\text{주변평균모형} : \text{logit}P(Y_{it} = 1|x_{it}) = x_{it}^T\beta, \tag{2.1}$$

$$\text{의존모형} : \text{logit}P(Y_{it} = 1|x_{it}, b_{it}) = \Delta_{it} + b_{it} \tag{2.2}$$

여기서 $b_i = (b_{i1}, \cdots, b_{in_i})^T \sim N(0, \Sigma_i)$, β는 주변평균모형의 $p \times 1$모수벡터이고, Δ는 절편이다. 여기서 Σ는 주로 AR(1)구조를 가지며 다음과 같다.

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n_i-1} \\ \rho & 1 & \rho & \cdots & \rho^{n_i-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n_i-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n_i-1} & \rho^{n_i-2} & \rho^{n_i-3} & \cdots & 1 \end{bmatrix} \tag{2.3}$$

위의 주변평균모형만을 가정하고 가상관계수행렬 (working correlation matrix)을 사용한 모형이 주변모형 (marginal model)이다. 의존모형에 있는 Δ_{it}대신에 $x_{it}^T\beta$로 교체한 것이 일반화선형혼합모형이다. 따라서 (2.1)-(2.3)에 주어진 주변화 변량효과모형은 주변모형과 일반화선형혼합모형의 모든 형태를 가지고 있음을 알 수 있다. 따라서 주변평균모형은 공변량의 모집단의 평균효과를 설명하는 모형이고, 의존모형 (2.2)는 반복된 결과치들의 상관관계를 설명하기 위한 의존모형이다. (2.1)과 (2.2)의 관계는 다음과 같은 관계가 있다.

$$P(Y_{it} = 1|x_{it}) = \int P(Y_{it} = 1|x_{it}, b_{it}) f(b_{it}) db_{it} \tag{2.4}$$

따라서 β와 σ^2이 주어지면 Δ_{it}는 (2.4)의 관계를 이용하여 뉴튼-라프슨 (Newton-Raphson) 알고리즘에 의해 계산된다. 모수(β, σ^2, ρ)들의 최대우도 추정치는 준뉴턴 (quasi-Newton) 알고리즘 이용하여 추정되면, 특히 β와 σ^2의 추정식 (estimating equation)에는 위의 (2.4)식의 관계에 의한 Δ_{it}의 편미분이 사용된다. 자세한 계산은 Heagerty (1999)에 제시되어 있다.

주변화모형은 일반화혼합모형과 비교하여 다음과 같은 장점들이 있다. 첫째, 주변평균모형 (2.1)를 통해서 한계효과 (marginal effect) 해석 시에 의존모형 (2.2)에 영향을 직접적으로 받지 않음을 알 수 있다. 따라서 일반화혼합모형과는 다르게 공변량의 직접적인 한계확률에 대한 효과를 설명할 수 있다. 둘째, GEE를 이용한 주변모형과는 다르게 의존모형 (2.2)를 통해서 우도함수를 구성할 수 있고, 이를 통하여 최대우도법이 가지는 다양한 장점들을 가질 수 있다. 예를 들면 최대우도추정치 (maximum likelihood estimate)의 일치성 (consistency)과 우도비검정 (likelihood ratio test)을 할 수 있다. 셋째, 만약 자료들이 결측치가 없으면서 변량효과변수의 분포가 틀리게 가정되었을 때, β의 추정시에 강건한 (robust) 추정치를 가짐을 알 수 있다 (Heagerty와 Zeger, 2000). 일반화혼합모형에서는 이 경우 편의가 발생할 수 있다 (Heagerty와 Kurland, 2001).

2.2. 변량효과 공분산행렬의 모형화

Heagerty (1999)에 제시한 주변화 변량효과모형인 (2.1)-(2.2)에서 변량효과공분산행렬은 (2.3)에 제시된 단순한 형태의 $AR(1)$구조를 가정하였다. 하지만 이러한 가정은 아주 강한 가정이며, 만약 임의결측치 (missing at random)가 있으면서 이 가정이 맞지 않을 때에는 β의 추정량에 편의가 발생함을 알 수 있다. 따라서 변량효과의 공분산행렬을 올바르게 추정할 필요성이 있고, 이를 위하여 Lee와 Sung (2014)는 변량효과의 공분산행렬에 MCD방법 (Pourahmadi, 1999)을 이용하여 이분산성과 $AR(p)$의 다양한 행태의 행렬을 일반화선형모형으로 모형화하였다. MCD방법은 현재의 변량효과 b_{it}를 그 이전의 변량효과 $b_{i1}, \cdots, b_{it-1}$를 이용하여 회귀식을 만든다. 그 형태는 다음과 같다.

$$b_{it} = \sum_{j=1}^{t-1} \phi_{i,tj} b_{ij} + e_{it} \tag{2.5}$$

(2.5)의 식을 벡터의 형태는 $T_i b_i = e_i$이고 여기서 T_i는 주대각 값들이 1이고 (t, j)위치 $(j < t)$의 값이 $-\phi_{i,tj}$인 하삼각행렬 (lower triangular matrix)이고 $e_i = (e_{i1}, \cdots, e_{in_i})^T$이다. 여기서 공분산을 계산하면 다음과 같이 된다.

$$T_i \Sigma_i T_i^T = D_i$$

여기서 $D_i = diag(\sigma_{i1}^2, \cdots, \sigma_{in_i}^2)$으로 e_i의 공분산행렬로 대각행렬 (diagonal matrix)이다. $\phi_{i,tj}$와 σ_{it}^2를 일반화자기상관모수 (generalized autoregressive parameter; GARP) 그리고 혁신변수 (innovation variance; IV)라고 한다. GARP와 IV를 일반화선형모형을 이용하여 변량효과공분산행렬의 이분산성과 일반적인 행태의 $AR(p)$구조를 설명할 수 있다. 그 형태는 다음과 같다.

$$\phi_{i,tj} = w_{i,tj}^T \gamma, \log(\sigma_{it}^2) = h_{it}^T \lambda \tag{2.6}$$

여기서 γ와 λ는 $a \times 1$과 $b \times 1$인 모수벡터이고, $w_{i,tj}$와 h_{it}는 GARP와 IV를 모형화하기 위해서 사용되는 디자인벡터이다. $w_{i,tj}$를 이용하여 AR구조의 변량효과공분산행렬을 만들 수 있다. 예를 들면 $w_{i,tj}^T = (I_{|t-j|=1})$는 $AR(1)$구조를 나타낸다. 여기서 $I_{|t-j|=1} = 1$만일 $|t-j| = 1$이면, 그 외는 0이다. 따라서

$$\phi_{i,tj} = w_{i,tj}\gamma = \begin{cases} \gamma, & \text{if } |t-j| = 1; \\ 0, & otherwise \end{cases}$$

이고,

$$b_{it} = \gamma b_{it-1} + e_{it}.$$

그리고 $h_{it}^T = (1, \text{sex}_i)$이면 변량효과공분산행렬이 성별에 따라 다른 공분산행렬을 가짐을 알 수 있다. 따라서 모형(2.6)에 의해서 GARP와 IV를 모형화한 변량효과공분산행렬은 다음의 장점들이 있다. 첫째, GARP와 log(IV)는 어떠한 제한도 없기 때문에 공변량을 이용하여 변량효과공분산행렬을 모형화할 수 있다. 둘째, h_{it}를 통하여 공분산행렬의 이분산성을 가질 수 있고, 그리고 $w_{i,tj}$를 통하여 일반적인 $AR(p)$구조의 공분산행렬을 가질 수 있다. 마지막으로 GARP와 IV를 일반화선형모형을 통하여 쉽게 모형화를 할 수 있게 된다.

3. 경시적 다범주자료 분석

Heagerty (1999)에 제안된 주변화 변량효과모형은 Lee와 Daniels (2008)과 Lee 등 (2011)에 의해서 각각 경시적 순서자료 (longitudinal ordinal data) 및 경시적 명목자료 (longitudinal nominal data)를 분석하는 모형으로 확장되었다. 이 절에서 이러한 확장을 고찰한다. 우선 자료의 형태는 다음과 같다. $Y_i^T = (Y_{i1}, \cdots, Y_{n_i})$를 i번째 개체의 K개의 범주를 가지는 순서자료의 응답변수벡터로 한다. 그리고 x_{it}는 i번째 개체의 t번째 독립변수로 $p \times 1$벡터로 한다.

3.1. 경시적 순서자료의 분석

우선 경시적 순서자료의 분석에 대한 주변화 변량효과모형을 보도록 하자. Lee와 Daniels (2008)에 의해서 제안된 이 모형은 Heagerty (1999)에 제시된 모형과 유사하다. 위의 Heagerty (1999)의 모형에서 $P(Y_{it} = 1|x_{it})$과 $P(Y_{it} = 1|x_{it}, b_{it})$을 대신하여 $P(Y_{it} \le k|x_{it})$와 $P(Y_{it} \le k|b_{it}, x_{it})$로 대치하여 누적로짓 (cumulative logit) 연결함수(link function)를 사용하였다. 여기서 $k = 1, \cdots, K-1$. 그 모형의 형태는 아래와 같다.

$$\text{주변평균모형} : \text{logit}P(Y_{it} \le k|x_{it}) = x_{it}^T\beta, \tag{3.1}$$

$$\text{의존모형} : \text{logit}P(Y_{it} \le k|x_{it}, b_{it}) = \Delta_{itk} + b_{it} \tag{3.2}$$

여기서 $b_i = (b_{i1}, \cdots, b_{in_i})^T \sim N(0, \Sigma_i)$로 Heagerty (1999)에 제시된 모형과 동일하다. 그리고 $\Delta_{it1} < \cdots < \Delta_{itK-1}$의 관계를 가지며 이 조건에 의해서 의존모형의 누적분포확률이 k에 따라서 증가성을 가짐을 알 수 있다. 위의 모형 (3.1)과 (3.2)은 Heagerty (1999)에 의해 제시된 모형과 유사하기 때문에 경시적 이진자료를 위한 주변화 변량효과모형과 동일한 장잠들을 가짐을 알 수 있다. 그리고 (3.1)과 (3.2)에 있는 누적분포확률의 관계에 의해 다음과 같은 식을 가진다.

$$P(Y_{it} \le 1|x_{it}) = \int P(Y_{it} \le 1|x_{it}, b_{it}) f(b_{it}) db_{it} \tag{3.3}$$

따라서 β와 σ^2이 주어지면 Δ_{it}는 (3.3)의 관계를 이용하여 뉴튼-라프슨 알고리즘에 의해 계산된다.
(3.2)에 있는 b_i의 공분산행렬인 Σ_i의 구조는 Lee와 Daniels (2008)에서 $AR(1)$을 가정하였다. 하지만 이러한 가정도 2.2에서 제시된 수정 콜레스키 방법을 이용하여 일반적인 $AR(p)$구조를 가지면서 이분산성을 가지는 변량효과공분산행렬로 모형화할 수 있다.

3.2. 경시적 명목자료 분석

Heagerty (1999)의 주변화 변량효과모형은 Lee 등 (2011)에 의해서 경시적 명목자료 (longitudinal nominal data)분석을 위해서 확장되었다. 여기서 제안된 모형은 다항로짓 (multinomial logit)을 가정하였다. 우선 Y_{itk}를 k범주 ($k = 1, \cdots, K-1$)를 나타내는 지시변수로 정의한다. 그러면 Lee 등 (2011)이 제안한 모형은 아래와 같다.

$$\log \frac{P(Y_{itk} = 1|x_{it})}{P(Y_{itK} = 1|x_{it})} = x_{it}^T\beta_k \tag{3.4}$$

$$\log \frac{P(Y_{itk} = 1|b_{it}, x_{it})}{P(Y_{itK} = 1|b_{it}, x_{it})} = \Delta_{itk} + b_{itk} \tag{3.5}$$

$$b_i \sim N(0, \Omega_i) \tag{3.6}$$

여기서 $b_i^T = (b_{i1}^T, \cdots, b_{in_i}^T) = (b_{i11}, \cdots, b_{i1K-1}, \cdots, b_{in_i1}, \cdots, b_{in_iK-1})$이고 β_k는 $p \times 1$인 회귀계수 벡터이다. 모형 (3.4)-(3.6) 역시 Heagerty (1999)의 모형을 확장한 모형이므로 그 장점과 특징을 공유한다.

변량효과 b_i의 공분산행렬인 Ω_i는 변량효과변수의 분산으로 크로레커곱 (Kronecker product)을 이용하여 시간에 따른 상관관계와 더불어 같은 시간에서의 범주들 간의 상관관계도 같이 고려한 모형을 가정하였다. 그래서 그 형태는 다음과 같다.

$$\Omega_i = R_i \otimes \Sigma$$

여기서 R_i는 식 (2.3)에서 제시된 상관계수행렬과 같은 형태이며, Σ는 범주들 간의 공분산행렬을 나타내고 그 행태는 다음과 같다.

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1K-1} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2K-1} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1K-1} & \sigma_{2K-1} & \cdots & \sigma_{K-1K-1} \end{bmatrix}$$

위의 식(3.4)-(3.6)로부터 다음의 관계를 가짐을 알 수 있다.

$$P(Y_{itk} = 1|x_{it}) = \int P(Y_{itk} = 1|b_{it}, x_{it}) f(b_{it}) db_{it}$$

위의 관계로 $\Delta_{it} = (\Delta_{it1}, \cdots, \Delta_{itK-1})$를 뉴튼-라프슨 알고리즘에 의해 계산됨을 알 수 있다.

4. 자료분석

4.1. 복지 패널자료 설명

본 논문에서 소개한 분석방법에 사용된 데이터는 한국복지패널연구로 수집하였다. 이 데이터는 2006년부터 2012년까지 가구별 설문조사를 하여 수집되었다 (http://www.koweps.re.kr). 총 7072가구를 대상으로 하여 가구당 복지에 관련된 설문으로 주거형태, 소득, 교육, 경제활동, 건강 및 의료, 사회보험/퇴직금제, 생활비 등을 조사한 자료이다. 본 데이터는 각 가구당 가구주만 추출한 데이터로 전체 객체 수는 7072명이고 이를 2006년부터 2012년까지 총 7번에 걸쳐 관찰 한 데이터 이다. 인구통계학적인 변수로는 각 가구당 할당된 ID (id), 가구별 가구주와의 관계 (pid), 성별 (0 = 여자, 1 = 남자), 나이 (15∼97), 교육수준 (1 = 중졸이하, 2 = 고졸, 3 = 대졸, 4 = 대학원이상), 결혼여부 (1 = 미혼, 2 = 기혼, 3 = 이혼, 사별, 별거 등등), 종교여부 (0 = 없음, 1 = 있음)으로 구성되어 있다.

4.2. 이진 복지패널자료 분석

4.1에서 제시된 자료를 분석하기 위해서 2.1에서 제시된 Heagerty (1999)의 주변화 변량효과모형을 이용하여 분석한다. 자료는 년도 별로 결측치가 발생하여 총 약 40%의 결측치가 존재한다. 2007에서 2012년에 걸쳐서 각각 10.8, 8.7, 4.8, 5.6, 6.5, 3.5%의 결측치가 존재한다, 이러한 결측치들을 임의 결측치 (missing at random)로 가정하고 분석하고자 한다. 응답변수로는 이진 값인 자동차 소유여부 (0

= 없음, 1 = 있음)이고 다음과 주변평균모형에 대해서 분석을 시행하였다.

$$\begin{aligned} \text{모형1} : \text{logit}P(Y_{it}=1|x_{it}) =& \beta_0+\beta_1 x_{\text{sex}}+\beta_2 x_{\text{age}}+\cdots+\beta_8 x_{\text{religion}} \\ \log(\sigma) =& \lambda_0+\lambda_1 \text{age} \end{aligned} \tag{4.1}$$

$$\begin{aligned} \text{모형2} : \text{logit}P(Y_{it}=1|x_{it}) =& \beta_0+\beta_1 x_{\text{sex}}+\beta_2 x_{\text{age}}+\cdots+\beta_8 x_{\text{religion}} \\ \log(\sigma) =& \lambda_0+\lambda_1 \text{sex} \end{aligned} \tag{4.2}$$

$$\begin{aligned} \text{모형3} : \text{logit}P(Y_{it}=1|x_{it}) =& \beta_0+\beta_1 x_{\text{sex}}+\beta_2 x_{\text{age}}+\cdots+\beta_8 x_{\text{religion}} \\ \log(\sigma) =& \lambda_0+\lambda_1 \text{age}+\lambda_2 \text{sex} \end{aligned} \tag{4.3}$$

위 모형식을 보면 주변평균모형은 모두 동일하지만 각 모형에 따른 $\log(\sigma)$가 다른 형태를 가짐으로써 공분산행렬을 모두 이분산성을 만족하면서 서로 다른 형태를 이루고 있음을 가정하였다. 여기서 age변수는 각 객체들의 나이를 log(age/10)로 변환한 변수이다. 따라서 모형1은 나이에 따라 이분산성을, 모형2는 성별로 이분산성을, 그리고 모형3은 나이와 성별에 따라서 이분산성을 가지는 변량효과공분산행렬을 가짐을 가정하였다.

Table 4.1 Comparison of maximized loglikelihoods, AICs, and BICs

	모형1	모형2	모형3
Max logL	-11105.38	-11115.63	-11105.11
AIC	22220.76	22241.26	22220.22
BIC	22254.79	22275.29	22254.25

위의 Table 4.1을 보면 AIC와 BIC 기준으로 봤을 땐, 모형1과 3의 AIC (BIC) 값이 22220.76 (22254.79), 22220.22 (22254.25)으로 모형2의 값보다 작은 값이 나온 것을 볼 수 있다. 그리고 모형1과 3에서는 모형의 단순성 (rule of parsimony)에 의해서 모형1을 가장 설명력이 있는 모형으로 선택한다. 각 적용방법에 대해서 추정한 변수값에 대한 결과는 표2에 주어져 있다.

Table 4.2 Maximum likelihood estimates and standard errors for three models

	모형1		모형2		모형3	
	estimate	std	estimate	std	estimate	std
Marginal mean ($\boldsymbol{\beta}$)						
Intercept	1.466^*	0.233	1.447^*	0.234	1.443^*	0.233
sex	1.404^*	0.076	1.409^*	0.076	1.407^*	0.076
log(age/10)	-2.346^*	0.120	-2.343^*	0.120	-2.336^*	0.119
edu2	0.706^*	0.071	0.723^*	0.072	0.712^*	0.072
edu3	1.175^*	0.091	1.185^*	0.093	1.181^*	0.090
edu4	1.419^*	0.119	1.412^*	0.124	1.422^*	0.119
marry1	-0.468^*	0.095	-0.435^*	0.090	-0.462^*	0.094
marry2	0.792^*	0.058	0.789^*	0.059	0.792^*	0.058
religion	0.031	0.025	0.035	0.026	0.032	0.025
$\log\sigma$						
Intercept	0.657^*	0.112	1.407^*	0.046	0.685^*	0.120
sex			-0.063	0.050	-0.032	0.050
log(age/10)	0.428^*	0.065			0.428^*	0.065
Max log L	-11105.38		-11115.63		-11105.11	
AIC	22220.76		22241.26		22220.22	
BIC	22254.79		22275.29		22254.25	

* indicates significance under 95% confidence level.

위 Table 4.2에 추정된 값들이 주어져 있다. 우선 $\log\sigma$의 추정치는 아래와 같다.

$$\log\widehat{\sigma}_i = 0.657 + 0.428\log(\text{age}_i/10).$$

여기서 log(age/10)의 계수의 추정치가 유의미함을 알 수 있다. 따라서 나이가 들어감에 따라 변량효과 분산이 더 커짐을 알 수 있다.

$$\begin{aligned}\log\frac{\widehat{P}(Y_{it}=1)}{\widehat{P}(Y_{it}=0)} =& 1.466 + 1.404\text{sex} - 2.346\text{age} + 0.706\text{edu2} \\ &+ 1.175\text{edu3} + 1.419\text{edu4} - 0.468\text{marry1} + 0.792\text{marry2} + 0.031\text{religion}\end{aligned}$$

성별이 남자 (sex=1)일 때 여자보다 차량의 소유의 로그오즈 (log odds)가 1.404만큼 증가 한 것을 알 수 있고, 나이가 더 많이 짐에 따라 차량의 소유 로그오즈가 2.346 만큼 감소함을 알 수 있다. 또한 고졸 (edu2=1)인 사람이 중졸이하 (edu=1)인 사람에 비해 로그오즈가 0.706 만큼 큼을 증가 한 것을 알 수 있고 같은 의미로 대졸 (edu3=1), 대학원이상 (edu4=1)은 각각의 로그오즈가 1.175, 1419 만큼 중졸이하인 사람들보다 증가한 것을 알 수 있다. 결혼여부도 마찬가지로 미혼 (marry1=1)인 사람이 이혼, 사별, 별거 등등 (marry=3)인 사람들에 비해 차량소유 로그오즈가 0.468 만큼 더 감소하였고 기혼 (marry2=1)인 사람의 로그오즈가 0.792 만큼 증가함을 알 수 있다.

5. 결론

우리가 모집단에서 응답변수의 평균에 대한 공변량의 효과에 관심이 있을 때, 주변모형 또는 주변화모형을 사용한다. 이 논문에서 우도함수에 기초한 주변화모형인 주변화 변량효과모형에 초점을 맞추었고, 이 모형이 측정된 결과치들의 상관관계를 변량효과 공분산행렬을 통하여 설명함을 보였다. 주변화 변량효과모형들은 주변평균모형과 의존모형으로 구성되며, 그 주변평균모형의 구성은 의존모형에 직접적으로 영향을 받지는 않기에 주변평균모형의 모수해석은 의존모형에 영향을 받지 않는다. 이러한 장점들로 인해 요즘 많은 자료 분석에서 주변화 변량효과모형을 이용하고 있고, 이 논문에서 다양한 형태의 경시적 범주형자료 분석에서 주변화 변량효과모형이 어떻게 쓰이는지를 살펴보았다.

본 논문에 사용된 복지 패널자료에 대한 분석결과를 보면 경시적자료에서 흔히 발생할 수 있는 임의결측치를 고려한 이분산성을 가지는 다양한 형태의 주변화 변량효과모형을 적합시켰고, 그 결과 나이와 성별에 따라서 공분산행렬을 다르게 가지는 모형이 가장 적합한 모형임을 알 수 있었다. 그리고 남자일수록, 나이가 젊을수록, 학력이 높을수록, 그리고 결혼경험이 있는 사람일수록 자동차소유에 대한 확률이 높게 나옴을 알 수 있었다.

References

Agresti, A. (2002). *Categorical data analysis*, 2nd edition, Wiley, New York.

Breslow, N. E. and Clayton, D. G. (1993). Approximate Inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 421, 9-25.

Daniels, M. J. and Hogan, J. W. (2008). *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*, Chapman & Hall/CRC, New York.

Diggle, P. J., Heagerty, P., Liang, K.-Y. and Zeger, S. (2002). *Analysis of longitudinal data*, 2nd Ed., Oxford Press, Oxford.

Heagerty, P. J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, **55**, 688-698.

Heagerty, P. J. (2002). Marginalized transition models and likelihood inference for longitudinal categorical data. *Biometrics*, **58**, 342-351.

Heagerty. P. J. and Kurland, B. F. (2001). Misspecified maximum likelihood estimate and generalized linear mixed models. *Biometrika*, **88**, 973-985.

Heagerty, P. J. and Zeger, S. L. (2000). Marginalized multilevel models and likelihood inference (with discussion). *Statistical Science*, **15**, 1-26.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal analysis using generalized linear models. *Biometrika*, **73**, 13-22.

Lee, K. and Daniels, M. J. (2007). A class of Markov models longitudinal ordinal data. *Biometrics*, **63**, 1060-1067.

Lee, K. and Daniels, M. J. (2008). Marginalized models for longitudinal ordinal data with application to quality of life studies. *Statistics in Medicine*, **27**, 4359-4380.

Lee, K., Kang, S., Liu, X. and Seo, D. (2011). Likelihood-based approach for analysis of longitudinal nominal data using marginalized random effects models. *Journal of Applied Statistics*. **38**, 1577-1590.

Lee, K. and Mercante, D. (2010). Longitudinal nominal data analysis using marginalized models. *Computational Statistics & Data Analysis*, **54**, 208-218

Lee. M., Lee. K. and Lee, J. (2012). Marginalized transition shared random effects models for longitudinal binary data with nonighorable dropout. *Biometrical Journal*, **56**, 230-242.

Lee. K. and Sung, S. (2013) Autoregressive Cholesky factor modeling for marginalized random effects models. *Communications for Statistical Applications and Methods*, **20**, 1-13.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Uunconstrained parameterisation. *Biometrika*, **86**, 677-690.

Journal of the Korean Data & Information Science Society
2014, **25**(6), 1263–1272

http://dx.doi.org/10.7465/jkdi.2014.25.6.1263
한국데이터정보과학회지

Review and discussion of marginalized random effects models[†]

Joo Yeong Jeon[1] · Keunbaik Lee[2]

[1]National Cancer Center
[2]Department of Statistics, Sungkyunkwan University

Received 29 July 2014, revised 1 September 2014, accepted 24 September 2014

Abstract

Longitudinal categorical data commonly occur from medical, health, and social sciences. In these data, the correlation of repeated outcomes is taken into account to explain the effects of covariates exactly. In this paper, we introduce marginalized random effects models that are used for the estimation of the population-averaged effects of covariates. We also review how these models have been developed. Real data analysis is presented using the marginalized random effects

Keywords: Conditional model, longitudinal data, marginal model, random effects, transition model.

[†] This project was supported by Basic Science Research Program through the National Research Foundation of Korea (KRF) funded by the Ministry of Education, Science and Technology (NRF-2012R1A1A1004002).
[1] Researcher, National Cancer Center, Seoul 410-769 Korea.
[2] Corresponding author: Associate professor, Department of Statistics, Sungkyunkwan University, Seoul 110-745, Korea. E-mail: keunbaik@skku.edu