응용통계연구 제 17권 2호, 2004년, pp. 337-346

집락자료의 분할표에서 독립성검정

정광모 [1] 이현영 [2]

요 약

랜덤표본에 관한 이원분할표의 독립성검정에는 통상 피어슨의 카이제곱적합도검정과 우도비검정을 사용한다. 그러나 랜덤표본이 아닌 집락자료에 관한 분할표의 경우에는 이들 검정법은 잘못된 결과를 나타낸다. 이러한 경우에는 공변량의 고정효과 외에 집락에 따른 변량효과를 함께 포함하는 일반화선형혼합모형을 고려함으로써 집락간의 이질성과 집락내의 종속성을 반영할 수 있다. 본 연구에서는 집락자료의 분할표에 대한 일반화선형혼합모형을 소개하고 실례를 통하여 이들 모형의 적합에 대해 논의한다.

주요용어: 집락자료, 독립성검정, 일반화선형혼합모형, 변량효과, 조건부로짓모형

1. 서론

이원분할표에서 행변수와 열변수간의 독립성검정에는 통상 피어슨의 카이제곱적합도검정을 사용한다. 그밖에 로그선형모형을 적합하여 우도비검정(likelihood ratio test: LRT)이나 왈드검정(Wald test)에 의해 교호작용효과를 가설검정할 수도 있다. 이 때 분할표는 랜덤표본에 의한 것이어야 하며 집락추출(cluster sampling) 표본에 관한 분할표의 경우에는 앞에서 언급된 검정통계량들은 적절하지 않다. 집락추출 표본의 경우 동일한 집락에 속하는 관찰값들은 서로 상관되어 있으므로 랜덤표본의 가정에 위배된다. 집락추출에 의한 자료, 동일한 개체에 대해 다시점에서 경시적으로(longitudinally) 반복측정한 자료, 인접한 공간영역에서 얻어지는 자료 등은 집락자료(clustered data)의 유형에 속한다. 예를 들어, 어미 쥐에서 출산한 같은 배의 새끼들은 서로 공통점을 갖기 때문에 이들에 관한 관찰값들은 상관이 높게 된다. 이러한 집락자료에 대한 모형이나 분석에서는 집락간의 이질성(heterogeneity)과 집락내의 종속성(dependency)을 반영하여야 한다. 다시점자료에 대한 통계적 모형에서는 전통적으로 일반화추정방정식(generalized estimating equation: GEE)을 이용하여 추정하게 되는데 GEE는 일차 및 이차 적률(moments)만으로 구해지기 때문에 우도함수를 전혀 이용하지 않는다는 비판이 있어 왔다. GEE에 관한 자세한 논의는 Agresti (2002)를 참고할 수 있다.

집락자료에서의 독립성검정에 관한 연구로 Altham (1976), Cohen (1976), Brier (1980), Rao 및 Scott (1981, 1984, 1987) 등은 피어슨 카이제곱검정이나 LRT를 수정한 검정법을 논의하였다. 반면에 Koch, Freeman 및 Freeman (1975), Binder(1983) 등은 왈드검정

1) (609-735) 부산시 금정구 장전동 산 30, 부산대학교 통계학과, 컴퓨터 및 정보통신연구소, 교수
E-mail: kmjung@pusan.ac.kr
2) (609-735) 부산시 금정구 장전동 산 30, 부산대학교 대학원 통계학과, 박사수료,
E-mail: dimes@pusan.ac.kr

에 근거한 검정법을 제안하였고, Fay(1985)는 피어슨통계량과 우도비검정통계량의 잭나이프(jackknife) 방법에 대해 논의하였다. 앞에서 언급한 여러 방법들은 Thomas, Singh 및 Roberts (1996)에 의해 종합적으로 개괄되었다.

공변량의 고정효과(fixed effect) 외에 집락에 대한 변량효과(random effect)를 함께 포함하는 혼합모형(mixed model)을 고려함으로써 집락간의 변이성과 집락내의 상관을 반영할 수 있다. 일반화선형모형 (generalized linear model:GLM)에 변량효과를 포함시킨 모형을 일반화선형혼합모형(generalized linear mixed model: GLMM)이라 한다. 분할표의 GLMM에 대한 연구로는 Brier(1980), Chowdhury와 McGilchrist(2001) 등이 있으며, 그밖에 Jain, Vilcassim 및 Chintagunta(1994), Revelt와 Train(1998) 등은 상품 선호도에 대한 반복측정 자료에 GLMM 을 적용하였다. GLMM은 반복측정 이항자료, 과산포(over dispersion) 이항분포 자료, 소지역별 자료, 소비자의 상품 선호도 조사자료, 사례-대조 비교실험을 하는 임상자료 및 집락추출에 의한 사회조사 자료 등에 널리 응용될 수 있다.

Lee와 Nelder(1996)는 GLMM에 대해 위계적우도(hierachical likelihood)를 최대로 하는 모수추정을 논의하였다. 그 후에 Chowdhury와 McGilchrist(2001)는 위계적우도와 수정된 LRT를 써서 집락추출에 대한 이원분할표의 독립성검정을 제안하였다. 반면에 주변우도(marginal likelihood)는 위계적 우도와 달리 변량효과가 주어진 조건하에서 관찰값의 조건부우도함수를 다시 랜덤효과의 분포함수에 관한 적분으로 나타내기 때문에 그 계산 절차가 매우 복잡하여 최대우도추정에 어려움이 따른다. 이러한 계산상의 문제를 해결하기 위해서는 통계소프웨어의 사용이 필수적이다. 본 연구에서는 GLMM을 적합시키는 몇 가지 절차를 구분하여 논의하고 실례를 통하여 추정 결과를 비교하고자 한다.

2. 집락자료에 대한 GLMM

집락추출 표본에 대한 분할표에서 집락 k의 i 번째 행과 j 번째 열에 해당하는 칸 도수를 y_{kij}라 하자. 집락 k 에서 어떤 개체가 cell (i,j)에 속하면 1, 그렇지 않으면 0을 나타내는 지시 확률변수 w_{kij} 의 평균을 $E(w_{kij}) = \mu_{kij}$라 하자. 분할표의 표본추출 모형을 포아송분포로 가정하면 w_{kij}는 평균 μ_{kij}인 포아송분포를 따르게 되고 이 때 칸 도수 y_{kij} 는 집락 k의 모든 개체에 관해 w_{kij}들을 합한 값으로 나타낼 수 있다. 분할표에서는 설명변수와 반응변수를 구분하지 않고 단지 두 변수간의 독립성이나 연관성(association)의 정도를 분석하지만 경우에 따라서는 이들 두 변수간의 역할을 구분하기도 한다. 이 경우 행변수는 설명변수, 열변수는 반응변수로 간주하여 앞의 경우와 다른 모형을 적합시키게 된다. 분할표에 대한 일반화선형모형은 크게 다항 로지스틱회귀모형(logistic regression model)과 포아송 로그선형모형(loglinear model)으로 구분할 수 있다. 전자의 모형은 반응변수와 설명변수를 구분하여 회귀모형을 적합하며, 후자의 모형은 두 변수간의 구분 없이 단지 연관성의 파악에 중점을 둔다. 로그선형모형을 변환하여 로지스틱회귀모형으로 표현할 수 있으며 또한 그 반대의 경우도 가능하기 때문에 이들은 서로 동치관계로 볼 수 있다. 칸 도수 y_{kij} 는 포아송분포 또는 다항분포를 따르는 확률변수로 간주할 수 있으나 앞으로 특별한 언급이 없으면 포아송분포를 가정하고 통계적 모형은 w_{kij} 의 평균 μ_{kij} 에 관해 나타내기로 하자.

집락효과를 고려하지 않은 경우 일반화선형모형은 단조증가인 연결함수(link function) $h(.)$ 를 써서

$$h(\mu_{kij}) = \mathbf{x}_{kij}{}'\boldsymbol{\beta} \tag{2.1}$$

와 같이 나타낼 수 있다. 식 (2.1)에서 $\mathbf{x}_{kij}$는 분할표에 관한 독립성모형이나 교호작용항을 포함하는 모형에 따라 정해지는 설계행렬(design matrix) 이고, $\boldsymbol{\beta}$는 이에 대응하는 모수들의 벡터이다. 연결함수 $h(.)$는 반응변수의 분포에 따라 선택하며 정준연결(canonical link)을 주로 사용한다. 예를 들어, 이항자료 및 포아송 도수자료에 대한 GLM의 정준연결은 각각 로지스틱연결과 로그연결이 된다. 정준연결은 이원분할표의 경우 상수항 θ_k, 행효과 α_i, 열효과 β_j 및 교호작용항 γ_{ij} 를 써서 모형 (2.1)의 우변을 $\theta_k + \alpha_i + \beta_j + \gamma_{ij}$ 와 같은 형태로 표현할 수 있으나 내용 전개의 일반성을 위해 식 (2.1)과 같이 나타내기로 한다.

모형 (2.1)은 집락효과를 고려하지 않은 경우이지만 만약 집락간의 이질성이나 집락내 관찰값들의 상관을 반영하기 위해서는 집락의 변량효과를 포함하는 혼합모형을 가정한다. 집락 k 의 변량효과를 $\mathbf{u}_k$ 라 하고 이의 분포함수를 $G(\mathbf{u}_k)$ 로 나타내자. 변량효과 $\mathbf{u}_k$ 와 고정효과 $\boldsymbol{\beta}$를 함께 포함하는 GLMM은

$$h(\mu_{kij}|\mathbf{u}_k) = \mathbf{x}_{kij}{}'\boldsymbol{\beta} + \mathbf{z}_{kij}{}'\mathbf{u}_k \tag{2.2}$$

와 같이 정의되며, 변량효과 $\mathbf{u}_k$는 평균 $\mathbf{0}$, 공분산 $\boldsymbol{\Sigma}$인 다변량정규분포를 따른다고 가정한다. 식(2.2)에서 $\mathbf{z}_{kij}$는 변량효과를 반영하는 설계행렬에 해당하며 변량효과의 지정에 따라 $\mathbf{z}_{kij}$가 정해진다. 식 (2.2)에서 연결함수 $h(.)$의 선택에 따라 분할표에 대한 대표적인 두가지 유형의 GLMM은 포아송 로그선형모형과 포아송 비선형모형이다.

1) 포아송 로그선형모형

포아송 평균 μ_{kij} 가 모수에 관해 선형식

$$\log(\mu_{kij}|\mathbf{u}_k) = \theta_k + \mathbf{x}_{kij}{}'\boldsymbol{\beta} + \mathbf{z}_{kij}{}'\mathbf{u}_k \tag{2.3}$$

으로 표현된다고 가정한다. 모형 (2.3)에서 $\mathbf{u}_k = (u_1, u_2, \cdots, u_J)$는 열변수의 범주에 대응하는 변량효과를 나타낸다. SAS 매크로인 GLIMMIX를 활용하면 유사우도함수에 관한 제한최대우도추정치(Restricted MLE)를 구하게 된다.

2) 포아송 비선형모형

로그선형모형 (2.3)대신에 포아송 평균 μ_{kij}가 아래와 같은 비선형 관계식

$$\mu_{kij} = \frac{\exp(\mathbf{x}_{kij}{}'\boldsymbol{\beta} + \mathbf{z}_{kij}{}'\mathbf{u}_k)}{\sum_i^I \sum_j^J \exp(\mathbf{x}_{kij}{}'\boldsymbol{\beta} + \mathbf{z}_{kij}{}'\mathbf{u}_k)} \tag{2.4}$$

으로 표현된다고 가정할 때, 이러한 모형을 포아송 비선형모형(Poisson nonlinear model)이라 한다. 참고로 집락의 총 관찰도수가 주어진 경우 y_{kij}는 μ_{kij}를 다항확률로 갖는 다항분포를 따른다고 간주할 수 있다. 모형 (2.4)는 모수에 관해 비선형인 점이 로그선형모형

(2.3)과의 차이점이다. SAS PROC NLMIXED는 이러한 비선형모형을 적합시키기 위해 고안된 프로그램이다.

그밖에 집락효과를 고려하지 않은 GLM 중 조건부로짓모형에 대해 간단히 살펴보자. 집락효과를 무시하는 경우에는 포아송 로그선형모형 (2.3)에서 변량효과 $\mathbf{u}_k$를 생략하여 고정효과만을 포함하는 모형을 적합시킨다. 분할표에서 행변수와 열변수를 각각 설명변수와 반응변수로 가정할 때 k번째 집락의 다항확률 π_{kij}에 대해 아래 식

$$\pi_{kij} = \frac{\exp(\mathbf{x}_{kij}'\boldsymbol{\beta})}{\sum_i^I \sum_j^J \exp(\mathbf{x}_{kij}'\boldsymbol{\beta})} \tag{2.5}$$

과 같이 가정하고 다항 로짓모형을 적합시킬 수 있다. 모형 (2.5)는 변량효과 $\mathbf{u}_k$와 절편항을 포함하지 않는다. McFadden (1973)은 이러한 모형을 조건부로짓모형 (conditional logit model)이라 하였으며, 좀더 일반적으로 이산선택모형(discrete choice model)이라 불리기도 한다. 조건부로짓모형은 소비자의 상품 선호도 자료의 모형화에 자주 이용된다. 다항반응 로짓회귀모형은 조건부로짓회귀모형의 특별한 유형에 속한다고 볼 수 있다. 이 모형의 적합에 생존분석에서 자주 인용되는 수명 자료에 대한 층화비례위험모형 (stratified proportional hazard model)의 적합 프로그램인 PHREG를 사용할 수 있는 이유는 조건부로짓모형에 대한 우도함수가 층화비례위험모형의 우도함수와 서로 동치가 되기 때문이다.

3. GLMM 적합

앞 절에서 소개한 두 가지 유형의 GLMM에 대해 모수추정을 위한 우도함수를 살펴보고 이들의 우도함수가 서로 동치임을 유도해 보자. 최대우도추정의 문제점과 근사계산법 및 통계소프트웨어의 특징에 대해서도 논의한다. 포아송로그선형모형 (2.3)을 가정할 때 $\mathbf{u}_k$가 주어진 경우 w_{kij}는 집락내에서 서로 독립이고, 또한 서로 다른 집락간에도 독립이라고 가정할 때 k번째 집락에서 w_{kij}의 로그우도는

$$\begin{aligned} & \sum_i^I \sum_j^J \log(\mu_{kij}^{w_{kij}} \exp(-\mu_{kij})) \\ \propto & \sum_i^I \sum_j^J w_{kij} \log(\mu_{kij}) - \sum_i^I \sum_j^J \mu_{kij} \\ = & \sum_i^I \sum_j^J w_{kij}(\theta_k + \mathbf{x}_{kij}'\boldsymbol{\beta} + \mathbf{z}_{kij}'\mathbf{u}_k) - 1 \end{aligned} \tag{3.1}$$

이다. 관계식

$$\sum_i^I \sum_j^J w_{kij} \exp(\theta_k + \mathbf{x}_{kij}'\boldsymbol{\beta} + \mathbf{z}_{kij}'\mathbf{u}_k) = \sum_i^I \sum_j^J w_{kij}\mu_{kij} = 1$$

에서

$$\theta_k = -\log(\sum_i^I \sum_j^J \exp(\mathbf{x}_{kij}'\boldsymbol{\beta} + \mathbf{z}_{kij}'\mathbf{u}_k)) \tag{3.2}$$

를 (3.1)식에 대입하면

$$\sum_i^I \sum_j^J w_{kij}(\mathbf{x}_{kij}'\boldsymbol{\beta} + \mathbf{z}_{kij}'\mathbf{u}_k) - \log(\sum_i^I \sum_j^J \exp(\mathbf{x}_{kij}'\boldsymbol{\beta} + \mathbf{z}_{kij}'\mathbf{u}_k)) - 1 \tag{3.3}$$

가 성립한다. 같은 요령으로 포아송 비선형모형 (2.4)에 대한 로그우도는

$$\begin{aligned} & \sum_i^I \sum_j^J w_{kij} \log(\mu_{kij}) - \sum_i^I \sum_j^J \mu_{kij} \\ \propto & \sum_i^I \sum_j^J w_{kij}(\mathbf{x}_{kij}'\boldsymbol{\beta} + \mathbf{z}_{kij}'\mathbf{u}_k) - \log(\sum_i^I \sum_j^J \exp(\mathbf{x}_{kij}'\boldsymbol{\beta} + \mathbf{z}_{kij}'\mathbf{u}_k)) - 1 \end{aligned} \tag{3.4}$$

이므로 식 (3.3)과 (3.4)에서 포아송 로그선형모형과 포아송 비선형모형에 대한 두 로그우도함수는 서로 동치가 됨을 알 수 있다.

앞에서 구해진 로그우도는 변량 $\mathbf{u}_k$가 주어질 때 조건부우도를 나타내며, 이 조건부우도를 $\mathbf{u}_k$의 분포 $G(\mathbf{u}_k)$에 관해 적분하면 주변우도가 구해진다. 편의상 $f(y_{ki1}, \cdots, y_{kiJ}|\mathbf{u}_k)$를 $\mathbf{u}_k$가 주어질 때 결합확률밀도라 하면 주변우도는 다음 식

$$\begin{aligned} L &= \int \prod\nolimits_k f(y_{ki1}, \cdots, y_{kiJ}|\mathbf{u}_k) dG(\mathbf{u}_k) \\ &= \int \prod\nolimits_k \prod\nolimits_{i,j} f(y_{kij}|\mathbf{u}_k) dG(\mathbf{u}_k) \end{aligned} \tag{3.5}$$

과 같다. 식 (3.5)의 주변우도는 다차원의 적분 계산이 포함되어 있기 때문에 모수의 추정과정이 매우 복잡하다.

주변우도의 적분 계산을 위해 Wolfinger 및 O'Connell(1993)은 유사우도(pseudo-likelihood: PL) 또는 제한유사우도(restricted PL)를 이용하여 수치적분 하였으며, Breslow 및 Clayton(1993)은 벌점준우도(penalized quasi-likelihood: PQL)를 제안하였다. PQL은 Wolfinger 및 O'Connell(1993)의 유사우도 알고리듬에 의해 구현될 수 있다. 그밖에 가우스-에르미트 구적법(Gauss-Hermite quadrature), 몬테칼로 EM 근사, 베이지안 MCMC 근사 계산 등이 있다. 가우스-에르미트 구적법 및 몬테칼로 EM 근사는 PQL 근사보다 MLE에 잘 수렴하지만 계산상 PQL 근사가 간단하다. 집락자료의 경우 PQL 근사에 의한 추정량은 편의(bias)를 갖는다. 편의를 줄이기 위해 집락내 관찰값들을 재추출(resampling)하여 모수를 추정하거나, 그밖에 붓스트랩(bootstrap)이나 잭나이프에 의해 편의를 줄이는 방법들이 있다. SAS 매크로 GLIMMIX는 Wolfinger 및 O'Connell(1993)의 PL 알고리듬에 의해 구현되었으며, 이것의 단점은 고정효과 및 공분산의 추정에 편의를 수반하는 점이다. 반면에 PROC NLMIXED는 가우스-에르미트 구적법을 사용하며 근사 계산이나 수치적분에 있어

서 GLIMMIX에 비해 우수한 특성을 갖는 것으로 알려져 있다. NLMIXED 및 GLIMMIX에서는 반응변수의 분포함수로 정규분포, 이항분포, 감마분포, 음이항분포, 포아송분포 등을 지정할 수 있다.

고정효과에 관한 가설 $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ 의 검정은 $\hat{\boldsymbol{\beta}}$가 점근적으로 카이제곱분포를 따른다는 사실을 이용한다. 즉, H_0이 참이고, 적당한 가정을 만족할 때 왈드통계량의 근사분포는

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)' \mathrm{I}^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \sim \chi^2 \tag{3.6}$$

을 만족한다. 단, I는 var($\hat{\boldsymbol{\beta}}$)의 일치추정량이다. 왈드통계량의 카이제곱 근사를 바탕으로 이와 동치인 우도비검정통계량

$$T_{lr} = 2[\, l(\hat{\boldsymbol{\beta}}, \phi) - l(\boldsymbol{\beta}_0, \phi)] \tag{3.7}$$

을 이용할 수 있다. 식 (3.7)에서 $l(\hat{\boldsymbol{\beta}}, \phi) = \log L$이고, ϕ는 산포(dispersion)를 나타낸다.

4. 실례를 통한 비교

[**보기 4.1**] 다음 자료는 Brier(1980)에서 인용한 것이다. 20개 마을(neighborhood)에 대해 각 마을에서 5가구씩 (어떤 마을은 3가구씩) 총 96가구를 추출하여 집에 대한 만족도를 3단계(불만족, 만족, 매우 만족) 순서척도로 측정하였다. 측정 변수는 이웃에 대한 만족도(변수 C)와 자기 집에 대한 만족도(변수 P)이고, 행 변수 C와 열 P변수 간의 독립성검정을 수행하고자 한다. 표 4.1에서 집락요인(마을)을 무시하면 변수 C와 변수 P에 관한 3×3 분할표로 요약할 수 있다.

집락추출에 의한 자료이기 때문에 한 마을에서 추출된 가구들은 서로 상관되어 있다. 따라서 피어슨의 카이제곱 독립성검정은 잘못된 결과를 나타낼 수 있다. 표 4.1의 자료에 관한 GLMM을 편의상 아래와 같이 나타내기로 하자. 마을의 변량효과를 u_k라 할 때 일반성을 잃지 않고 상수항 θ_k를 0으로 가정하면 이원분할표에 대한 GLMM은

$$h(\mu_{kij} | \mathbf{u}_k) = \alpha_i + \beta_j + \gamma_{ij} + \mathbf{z}_{kij}' \mathbf{u}_k \tag{4.1}$$

와 같다. 행 변수를 나타내는 변수 C를 설명변수 x_{kij}라 하자. 행효과 α_i를 설명변수에 흡수시키고 열범주에 따른 변량효과만을 고려하게 되면 모형 (2.2)는 간단히

$$h(\mu_{kij} | \mathbf{u}_k) = x_{kij}\beta_j + \gamma_{ij} + u_{kj} \tag{4.2}$$

와 같이 나타낼 수 있다. 단, $k = 1, \cdots, 20$이고, $i, j = 1, 2, 3$이다. 모수 β_j는 열변수 P의 효과, γ_{ij}는 행변수 C와 열변수 P의 교호작용 효과를 나타낸다. 모수 β_j를 추정할 때 맨 끝 범주에 대응하는 β_3는 통상 0으로 가정하기 때문에 실제로는 β_1, β_2 두 개만을 추정하면 된다. 같은 요령으로 교호작용항은 총 네 개를 추정하게 된다. 여기서 u_{kj}는 j번째 열범주에 대한 변량효과로서 정규분포를 따른다고 가정한다. 만약 모든 i, j에 대해 $\gamma_{ij} = 0$이면 독립성모형(independence model)이 되므로 두 변수간의 독립성검정에 관한 통계적가설은

$$H_0 : \gamma_{ij} = 0, i = 1, 2, 3; j = 1, 2, 3 \tag{4.3}$$

와 같이 나타낼 수 있다. 따라서 독립성모형과 연관성모형에 대한 자유도 차이는 4가 된다.

표 4.1: 20개 마을에서 조사된 집에 대한 만족도

마을	C_1P_1	C_1P_2	C_1P_3	C_2P_1	C_2P_2	C_2P_3	C_3P_1	C_3P_2	C_3P_3	합계
1	1	0	0	2	2	0	0	0	0	5
2	1	0	0	2	2	0	0	0	0	5
3	0	2	0	0	2	0	0	1	0	5
4	0	1	0	2	1	0	1	0	0	5
5	0	0	0	0	4	0	0	1	0	5
6	1	0	0	3	1	0	0	0	0	5
7	3	0	0	0	1	0	0	1	0	5
8	1	0	0	1	3	0	0	0	0	5
9	3	0	0	0	0	0	1	0	1	5
10	0	1	0	0	3	1	0	0	0	5
11	1	1	0	0	2	0	1	0	0	5
12	0	1	0	4	0	0	0	0	0	5
13	0	0	0	4	1	0	0	0	0	5
14	0	0	0	1	2	0	0	0	2	5
15	2	0	0	2	1	0	0	0	0	5
16	1	0	0	1	1	0	0	0	0	3
17	0	0	0	1	1	1	0	2	0	5
18	0	0	0	1	0	1	0	0	1	3
19	2	0	0	2	1	0	0	0	0	5
20	2	0	0	2	0	0	1	0	0	5
총계										96

이웃에 대한 만족도와 자기집에 대한 만족도가 서로 독립이라는 귀무가설에 대해 SAS에서 구현할 수 있는 대표적인 프로그램들에 의한 수행 결과는 표 4.2와 같다. 그밖에 PROC GENMOD에 REPEATED 명령어를 사용하여 독립성모형과 교호작용모형 각각에 대해 GEE적합을 실행한 결과 독립성모형의 경우에는 표 4.2의 GENMOD와 동일한 추정치가 구해졌으나 교호작용모형의 경우에는 추정치가 수렴하지 않았다. PROC GENMOD와 PHREG는 집락에 대한 변량효과를 고려하지 않은 모형이다. 표 4.2에서 네 가지 프로그램에 의한 추정치의 크기는 $\hat{\beta}_1 > \hat{\beta}_2 > \hat{\beta}_3$로서 조사 대상자들은 대체로 자기 집에 불만족하는 경향을 나타낸다. PHREG에 의한 추정치가 가장 큰 반면 NLMIXED는 가장 작은 추정치를 나타내고, GENMOD와 GLIMMIX에 의한 추정치는 매우 유사하다. 우도비통계량의 유의성에 있어서는 GENMOD가 변량효과를 고려한 NLMIXED나 GLIMMIX에 비해 훨씬 더 유의한 결과를 나타내고 있어 집락자료의 경우 변량효과를 고려하지 않은 GENMOD의 사용에 특별한 주의를 요한다. 또한 변량효과를 포함하지 않는 조건부로짓모형에 대한 PHREG의 유의성검정 결과가 NLMIXED와 매우 유사하여 흥미있는 결과를 보여주고 있

표 4.2: 독립성모형에 대한 모수추정치 및 우도비검정결과

추정치	GENMOD	PHREG	GLIMMIX	NLMIXED
$\hat{\beta}_1$	1.966(0.4036)	2.465(0.4317)	1.943_1	1.658(0.3858)
$\hat{\beta}_2$	1.718(0.4015)	2.046(0.3917)	1.688(0.3194)	1.476(0.3919)
분산추정치	-	-	$\hat{\sigma}_{11} = 0.1584$ $\hat{\sigma}_{22} = 0.0455$ $\hat{\sigma}^2 = 0.5761$	$\hat{\sigma}_{11} = 1E-8$ $\hat{\sigma}_{22} = 1E-8$
우도비통계량	15.39	11.19	12.96	10.9
(유의성)	$(0.004)_2$	(0.0245)	(0.0115)	(0.0277)

[1] 가변수를 사용하였기 때문에 추정치에 상응하는 표준오차가 출력되지 않음

[2] 피어슨통계량은 $X^2 = 17.90$이고 이때 P=0.0013임

다. 일반적으로 집락추출 자료에 대해 변량효과를 포함하지 않게 되면 추정량의 표준오차가 실제보다 작게 추정되는 경향이 있으며 따라서 가설검정의 유의성이 GLMM 적합에 비해 높게 된다.

5. 맺음말

본 연구에서는 집락추출 표본에 관한 분할표에서 행과 열 변수간의 독립성검정에 대해 알아보았다. 통상적인 피어슨 카이제곱검정은 집락효과를 전혀 반영하지 못하기 때문에 이것을 그대로 사용하게 되면 검정통계량의 유의성이 실제보다 커지는 문제점이 있다. 집락효과를 반영한 변량효과모형 중 포아송 로그선형모형과 포아송 비선형모형의 적합에 이용할 수 있는 SAS 매크로 GLIMMIX와 PROC NLMIXED 프로그램의 모수추정 및 그 특징에 대해 살펴보았다. 실례를 통해 독립성검정에 대한 우도비검정통계량의 관점에서 이들 프로그램에 의한 유의성을 비교하였다. 집락추출에 의한 관찰값의 이원분할표에서 그밖에 PROC PHREG에 의해 변량효과를 갖지 않는 조건부로짓모형을 적합시킬 수 있음을 논의하였다. 각각의 프로그램을 적용하기 위해서는 입력 자료가 그에 맞는 적절한 형태로 구성되어야 한다.

집락자료의 대표적인 모형적합 방법인 GLMM, GEE 및 조건부로짓모형에 관해 추정량의 편의(bias), 표준오차 및 효율성 등이 Allison(2001)에서 언급된바 있는 데 GLMM과 GEE가 선호된다. 그러나 이러한 방법들은 실험계획이나 표본설계, 자료 및 분석 목적에 따라 영향을 받는다. 예를 들어, 조건부로짓모형은 실험자료의 경우에는 별로 바람직하지 못하지만 관찰자료(observational data)의 경우에는 매우 매력적인 것으로 알려져 있다.

순서반응(ordinal response)을 고려한 다항로짓모형, 여러 개 집락요인이 있는 집락추출 표본에 대한 분할표의 분석 등은 앞으로의 연구과제로 남겨둔다.

참고문헌

Agresti, A. (2002). *Categorical Data Analysis*, 2nd ed.,Wiley.

Allison, P.D. (2001). *Logistic Regression,Using the SAS System,* Theory and Application, SAS Institute Inc.

Altham, P.M.E. (1976). Discrete variable analysis for individuals grouped into families, *Biometrika*, **63**, 263-269.

Binder, D. A. (1983). On the variance of asymptotically normal estimators from complex surveys, *International Statistical Review*, **51**, 270-292.

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

Brier, S. S. (1980). Analysis of contingency tables under cluster sampling, *Biometrika*, **67**, 591-596.

Chen, Z. and Kuo, L. (2001). A note on the estimation of the multinomial logit model with random effects, *The American Statistician*, **55**, 89-95

Chowdhury, S. R. and McGilchrist, C. A. (2001). Analysis of contingency tables with clustered observations, *Australian and New Zealand Journal of Statistics*, **43**, 351-358.

Cohen, J.E. (1976). The distribution of the chi-squared statistic under cluster sampling from contingency tables, *Journal of the American Statistical Association*, **75**, 261-268.

Fay, R. E. (1985). A Jackknifed chi-squared tests for complex samples, *Journal of the American Statistical Association*, **80**, 51-60.

Jain, D. C., Vilcassim, N. J. and Chintagunta, P. K. (1994). A random-coefficient logit Brand-Choice model applied to panel data, *Journal of Business and Economic Statistics*, **12**, 317-328.

Koch, G. G., Freeman, D. H. and Freeman, I. J. (1975). Strategies in the multivariate analysis of data from complex surveys, *International Statistical Review*, **43**, 59-78.

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

McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behaviour, in *Frontiers of Econometrics*, ed. P. Zarembka, New York: Academic Press, 105-142.

Rao, J. N. K. and Scott, A. J.(1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables, *Journal of the American statistical Association*, **76**, 221-230.

Rao, J. N. K. and Scott, A. J.(1987). On simple adjustments of chi-squared tests with sample survey data, *The Annals of Statistics*, **15**, 385-397.

Revelt, D. and Train, K. (1998). Mixed logit with repeated choices: Households' choices of appliance efficiency level, *The Review of Economics and Statistics*, **80**, 647-657.

Thomas, D. R., Singh, A. C. and Roberts, G. R. (1996). Test of independence on two-way tables under cluster sampling : *An Evaluation, International Statistical Review*, **64**, 295-311.

Wolfinger, R. and O'Connell, M.(1993). Generalized linear mixed models: a pseudo-likelihood approach, *Journal of Statistical and Computational Simulation*, **48**, 233-243.

[2003년 9월 접수, 2003년 11월 채택]

Testing Independence in Contingency Tables with Clustered Data

Kwang Mo Jeong [1] Hyun Young Lee [2]

ABSTRACT

The Pearson chi-square goodness-of-fit test and the likelihood ratio tests are usually used for testing independence in two-way contingency tables under random sampling. But both of these tests may provide false results for the contingency table with clustered observations. In this case we consider the generalized linear mixed model which includes random effects of clustering in addition to the fixed effects of covariates. Both the heterogeneity between clusters and the dependency within a cluster can be explained via generalized linear mixed model. In this paper we introduce several types of generalized linear mixed model for testing independence in contingency tables with clustered observations. We also discuss the fitting of these models through a real dataset.

Keywords: Clustered data, Independence test, Random effect, Generalized linear mixed model, Conditional logit model.

1) Professor, Department of Statistics, Research Institute of Computer, Information and Communication, Pusan National University, Jangjeon-dong 30, Pusan, 609-735, Korea
E-mail: kmjung@pusan.ac.kr

2) Graduate, Department of Statistics, Pusan National University, Jangjeon-dong 30, Pusan, 609-735, Korea
E-mail: dimes@pusan.ac.kr