응용통계연구 제 **15**권 *1*호, *2002*년, *pp. 129-137*

다가자료에 대한 혼합효과모형 *

최재성[1)]

요약

본 논문은 개체의 반응에 영향을 미치는 독립변수들중 일부는 고정요인들이고 일부는 확률요인들로 간주되며 반응변수가 다가범주를 갖는 명목형 변수일 때, 다원분류표에서 자료를 분석하기 위한 모형으로 혼합효과모형을 제시하고 모형내 미지모수들을 추정하는 방법을 다루고 있다.

주요용어: 다가자료, 기준범주 로짓, 혼합효과.

1. 서론

범주형 자료를 개체 또는 실험단위의 반응에 대한 관측범주들의 수로 분류할 때, 이가자료(binary data)와 다가자료(polytomous data)로 구분할 수 있다. 이가자료는 개체들의 반응변수에 대한 관측값들이 두개의 범주만으로 관측되는 자료를 말한다. 다가자료는 개체들의 반응에 대한 관측값들이 셋이상의 유한개의 범주들로 구성되는 자료를 의미한다. 이가자료가 개체의 두 반응범주의 도수로 주어질 때 이가자료는 이항자료(binomial data)를 이루게 된다. 이들 이가자료 및 집단화된 이가자료(grouped binary data) 또는 이항자료에 대한 자료분석방법들은 Cox and Snell(1989)이 다양한 경우에 대해 논의하고 있다. 다가자료는 자료의 특성상 이가자료 또는 이항자료와는 달리 자료구조의 복잡성때문에 모형에 근거한 분석방법도 용이하지 않음을 예상할 수 있다. 다가자료의 구조적 특성을 고려한 다양한 모형들 및 분석방법들은 McCullagh and Nelder(1989) 와 Agresti(1990)에서 논의되고 있다. 본 논문은 개체의 반응이 셋 이상의 다가범주로 주어지는 명목형의 반응변수이고 반응에 영향을 미치는 독립변수들의 일부가 고정효과를 나타내는 설명변수들이며 다른 변수들은 확률효과를 나타내는 설명변수들일 때, 다가자료를 분석하기 위한 혼합효과 모형을 고려하고 있다. 이항자료의 초과변동(overdispersion)을 다루기 위한 혼합효과 모형에 관한 논의는 Williams(1982a)에서 살펴볼 수 있다. Im and Gianola(1988)는 이원지분계획으로 부터 발생하는 분산성분들을 추정하기 위하여 이항자료에 대한 혼합효과모형을 이용하고 있으나 다가자료의 혼합효과 모형에 관한 연구는 문헌에서 인용하기가 쉽지 않다. 개체 또는 실험단위의 반응에 대한 단순척도의(pure scale) 관측범주들이 다가의 범주로 주어질 때, 실험 또는 조사로 부터 수집된 자료는 다가자료를 구성하게 되고 표본의 크기가 주어진 경우에 이들자료가 반응범주의 도수로 표현되면 다항자료(multinomial data)라 불리어 진다. 본 연구는 관심모집단의 개체에 대한 다수의 특성들중 한 변수가 반응변수이고 다른 변수들은 반응에 영향을 미치는 설명변수들일 때, 이중 일부변수들의 수준들은 고정되어 있는

1) 본 연구는 2000년도 계명대학교 비사연구기금으로 이루어졌음
1) (704-701) 대구광역시 달서구 신당동 1000, 계명대학교 통계학과 교수

요인들로 간주되고 다른변수들의 수준은 확률표본으로 취해진다고 가정한다. 따라서 반응변수와 설명변수들의 관계를 규명하기 위한 분석모형은 고정요인들의 효과와 확률요인들의 효과를 포함하고 있는 혼합효과 모형이 된다. 혼합효과 모형내 모수를 추론하기 위해 이용되는 자료는 설명변수들의 각 수준결합에서 독립적인 다항분포를 갖는 다항자료가 된다. 주어진 모형하의 모수에 관한 추론은 구체적인 예를 통하여 논의하고자 한다.

2. 모형

관심모집단내 개체에 대한 반응이 명목형의 다범주(multi-category)로 주어지고 각 반응범주의 확률에 영향을 미치는 독립변수들로 두 요인 A와 B를 고려한다. 요인 A는 $i = 1, 2, ..., a$ 개의 수준들로 이루어진 고정요인(fixed factor)이고 요인 B는 수준들의 집단에서 임의로 추출된 $j = 1, 2, ..., b$ 개의 수준들로 구성된 확률요인(random factor)이라 가정한다. 개체에 대한 반응은 $h = 1, 2, ..., c$ 개의 범주들로 주어지는 반응변수 Y로 나타낸다. 연구자의 관심 모집단에서 개체의 세가지 특성 A, B와 Y에 대한 조사는 두 요인들의 수준결합에서 Y의 값들에 대한 조건부 확률들의 분포로 주어진다. 즉, $\{\pi_{h|ij}\}$ 이다.반응변수 Y의 관측값들로 나타나는 반응범주들의 확률에 영향을 미치는 두 요인들의 수준효과를 추론하기 위하여, 두 요인들의 수준결합에서 크기 n_{ij} 인 독립인 확률표본을 취한다고 하자. 여기서 n_{ij}는 요인 A의 i 번째 수준과 요인 B의 j번째 수준에서 취해진 확률표본이다. 표본크기가 n_{ij} 인 요인 A의 i 번째 수준과 요인 B의 j번째 수준에서 반응변수 Y의 c개 범주들의 관측도수를 n_{ijh} 로 나타낼 때, n_{ijh} 들은 다항분포를 따른다. 이들 자료를 분석하기 위한 모형은 두 요인들을 측정척도에 따른 범주형 변수들로 구분할 때 다수의 가능한 모형들을 고려할 수 있다. 첫 째는 두 요인들이 모두 명목형 변수들일 때, 자료분석 모형은 다음과 같다.

$$g(P(Y = h|ij)) = \alpha_h + {\beta_{ih}}^A + {\beta_{jh}}^B \tag{2.1}$$
$$i = 1, ..., a, j = 1, ..., b, h = 1, ..., c - 1.$$

여기서 $g(.)$는 연결함수이고, α_h는 반응변수 Y가 범주 h로 반응할 때의 절편을 나타내며 $\{{\beta_{ih}}^A\}$ 는 요인 A의 고정효과를 $\{{\beta_{jh}}^B\}$는 요인 B의 확률효과들을 나타낸다. 또한 $\{{\beta_{jh}}^B\}$는 확률효과들이므로 $N(0, {\sigma_h}^2)$을 따른다고 가정한다.두번 째는 요인 A가 명목형 변수이고 요인 B가 순서형 변수일 때의 모형이다.

$$g(P(Y = h|ij)) = \alpha_h + {\beta_{ih}}^A + v_j\tau_h \tag{2.2}$$
$$i = 1, ..., a, j = 1, ..., b, h = 1, ..., c - 1.$$

단, $\{v_j\}$는 요인 B의 수준들과 동일한 순서를 갖는 단조점수들이고 $\{\tau_h\}$는 연결함수 g로 변환된 값들에 대하여 반응범주들에 따른 기울기를 나타낸다.세번 째는 두 요인들이 모두 순서형 변수들일 때의 모형을 고려할 수 있다.

$$g(P(Y = h|ij)) = \alpha_h + u_i\theta_h + v_j\tau_h \tag{2.3}$$
$$i = 1, ..., a, j = 1, ..., b, h = 1, ..., c - 1.$$

단, $\{u_i\}$는 요인 A의 수준들과 동일한 순서를 갖는 단조점수들이고 $\{\theta_h\}$는 연결함수g 로 변환된 값들에 대하여 반응범주들에 따른 기울기를 나타낸다.세 변수에 대한 삼원분할표의 자료를 분석하기 위한 포화모형(saturated model)은 다음과 같다.

$$g(P(Y=h|ij)) = \alpha_h + \beta_{ih}{}^A + \beta_{jh}{}^B + \gamma_{ijh}$$

단, $\{\gamma_{ijh}\}$는 두 요인 A와 B의 교호작용을 나타낸다.반응변수가 하나 존재하는 삼원분할표의 도수자료를 분석하기 위한 모형은 설명변수의 유형에 따라서 다양하게 주어짐을 알 수 있다. 이 경우에 이용되는 연결함수는 주로 로짓연결함수이다. 반응변수가 셋 이상의 다범주를 갖는 명목형 변수일 때 일반화 로짓은 기준범주 로짓(baseline-category logits)을 이용하게 된다. 모형(1)에 해당하는 일반화 로짓모형(generalized logit model)을 나타내기 위하여 기준범주를 c로 택한다.$\pi_{h|ij}$를 요인 A가 수준 i이고 요인 B가 수준 j일 때 개체에 대한 반응이 범주 h로 관측될 조건부 확률이라 두자. 이들 기준범주 로짓을 이용한 일반화 로짓 혼합효과모형은 다음과 같이 표현된다.

$$log(\frac{\pi_{h|ij}}{\pi_{c|ij}}) = \alpha_h + \beta_{ih}{}^A + \beta_{jh}{}^B \tag{2.4}$$
$$i = 1, ..., a, j = 1, ..., b, h = 1, ..., c-1.$$

3. 모수의 추론

모형내 모수들의 추론을 위하여 요인 A의 수준 i와 요인 B의 수준 j에서 크기 n_{ij}인 확률표본을 취한다고 가정하자. 이때 $\{n_{ijh}\}$들은 c개의 반응범주에 속할 확률이 각각 $\pi_{1|ij}$, $\pi_{2|ij}$, ..., $\pi_{c|ij}$인 다항분포를 따르게 된다. 두 요인들의 서로 다른 수준결합에서 취해지는 확률표본은 독립적으로 행해지기 때문에 관측도수들의 확률분포로써 독립적인 다항분포를 가정하게 된다.반응변수의 범주가 c개이므로 c-1 개의 일반화 로짓모형들을 고려해야한다. 즉, 모형(5)를 이용할 때 c-1 개의 방정식들은 다음과 같다.

$$\begin{aligned}
log(\frac{\pi_{1|ij}}{\pi_{c|ij}}) &= \alpha_1 + \beta_{i1}{}^A + \beta_{j1}{}^B \\
log(\frac{\pi_{2|ij}}{\pi_{c|ij}}) &= \alpha_2 + \beta_{i2}{}^A + \beta_{j2}{}^B \\
&\cdot \\
&\cdot \\
&\cdot \\
log(\frac{\pi_{c-1|ij}}{\pi_{c|ij}}) &= \alpha_{c-1} + \beta_{i,c-1}{}^A + \beta_{j,c-1}{}^B \\
i &= 1, ..., a, j = 1, ..., b.
\end{aligned} \tag{3.1}$$

모형(5)는 두 요인 A와 B가 둘다 명목형의 고정요인들인 경우에 이들의 효과들을 추론하기 위한 모형이다. 한편, 요인 A는 명목형 고정요인이고 요인 B는 확률요인이라 가정할 때 요인 B의 수준들은 일반적으로 정규분포하는 한 확률변수의 값들로 간주된다. 이때, 관심

모수는 이들 수준의 개별적인 효과에 있다기 보다는 이들 효과의 변이정도에 관심을 갖게 된다. 따라서, 요인 B의 확률효과를 고려한 모형은 ${\beta_{jh}}^B = \sigma_h z_j$로 둘때 위의 방정식들은 다음과 같이 표현된다.

$$\begin{aligned} log(\tfrac{\pi_{1|ij}}{\pi_{c|ij}}) &= \alpha_1 + {\beta_{i1}}^A + \sigma_h z_j \\ log(\tfrac{\pi_{2|ij}}{\pi_{c|ij}}) &= \alpha_2 + {\beta_{i2}}^A + \sigma_h z_j \\ &\cdot \\ &\cdot \\ &\cdot \\ log(\tfrac{\pi_{c-1|ij}}{\pi_{c|ij}}) &= \alpha_{c-1} + {\beta_{i,c-1}}^A + \sigma_h z_j \\ i &= 1,...,a, j = 1,...,b. \end{aligned} \tag{3.2}$$

단, z_j는 평균이 0이고 분산이 1인 표준정규변수이다.모형내 모수들의 추정을 위하여 두 요인의 모든 수준결합에서 관측되는 도수들의 결합분포는

$$\begin{aligned} &\textstyle\prod_{i=1}^a \prod_{j=1}^b \{\frac{n_{ij+}!}{n_{ij1}!n_{ij2}!...n_{ijc}!} \\ &{\pi_{1|ij}}^{n_{ij1}} {\pi_{2|ij}}^{n_{ij2}} ... {\pi_{c|ij}}^{n_{ijc}}\} \end{aligned} \tag{3.3}$$

로 주어진다. 여기서 n_{ij+}는 요인 A의 수준 i와 요인 B의 수준 j에서 반응변수 Y의 c개 범주에서 관측되는 도수들의 합을 나타낸다.식(7)에 모형(6)를 대입하여 모형내 모수들의 함수인 다음과 같은 조건부 우도함수를 얻게 된다. 확률벡타 $\boldsymbol{Z} = \boldsymbol{z}$가 주어졌을 때 조건부 우도함수를 $\boldsymbol{L(\alpha, \beta, \sigma^2, z)}$라 두자.

$$\begin{aligned} \boldsymbol{L(\alpha, \beta, \sigma^2, z)} \quad &= \textstyle\prod_{i=1}^a \prod_{j=1}^b \{\frac{n_{ij+}!}{n_{ij1}!n_{ij2}!...n_{ijc}!}\} [\frac{exp(\alpha_1 + {\beta_{i1}}^A + \sigma_1 z_j)}{\sum_{s=1}^c exp(\alpha_s + {\beta_{is}}^A + \sigma_s z_j)}]^{n_{ij1}} \\ &[\tfrac{exp(\alpha_2 + {\beta_{i2}}^A + \sigma_2 z_j)}{\sum_{s=1}^c exp(\alpha_s + {\beta_{is}}^A + \sigma_s z_j)}]^{n_{ij2}} \cdots [\tfrac{exp(\alpha_{c-1} + {\beta_{i,c-1}}^A + \sigma_{c-1} z_j)}{\sum_{s=1}^c exp(\alpha_s + {\beta_{is}}^A + \sigma_s z_j)}]^{n_{ijc}} \end{aligned} \tag{3.4}$$

단, $\boldsymbol{\alpha}$= $(\alpha_1, \alpha_2, ..., \alpha_{c-1})'$ 이고 $\boldsymbol{\beta}$=$(\boldsymbol{{\beta_1}^A, {\beta_2}^A, ..., {\beta_{c-1}}^A})'$ 이며,$\boldsymbol{{\beta_h}^A}$= $({\beta_{1h}}^A, {\beta_{2h}}^A, ..., {\beta_{ah}}^A)$ 이다. $\boldsymbol{\sigma^2}$= $({\sigma_1}^2, {\sigma_2}^2, ..., {\sigma_{c-1}}^2)'$, $\boldsymbol{z}$= $(z_1, z_2, ..., z_b)'$ 이며, $h = 1, 2, \cdots, c-1$이다. 조건부 우도함수 $\boldsymbol{L(\alpha, \beta, \sigma^2, z)}$은 미지의 모수벡타 $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$와 $\boldsymbol{\sigma^2}$에 의존하며 또한 확률벡타 $\boldsymbol{z}$의 함수이다. 미지모수들을 추정하기 위하여 $\boldsymbol{z_j}$들은 서로 독립인 표준정규분포를 따른다고 가정한다. 구체적인 확률분포를 갖는 확률변수들을 포함하고 있는 조건부 우도함수를 다루는 표준적인 방법은 이들 변수들의 분포에 대하여 조건부 우도함수를 적분하는 것이다. 이들 확률변수에 대해 적분한 후, 얻어지는 결과의 함수는 주변우도함수(marginal likelihood function)이며 이 함수는 미지모수들 만의 함수이다. 모수들의 최우추정치는 이 주변우도함수를 최대로 하는 값들이다. 주변우도함수를 이용하여 모형내 모수들의 최우추정치들을 구하기 위하여 우선 조건부 우도함수 $\boldsymbol{L(\alpha, \beta, \sigma^2, z)}$를 b개의 $\boldsymbol{z_j}$들에 대하여 적분한 후 대수변환하여 얻은 결과의 함수는 주변대수우도함수(marginal log-likelihood function)이다. 이 함수를 MLLH라 둘 때,

$$\boldsymbol{MLLH = log\{\int \cdots \int L\{\prod_{j=1}^b \frac{1}{\sqrt{2\pi}} \exp\Big\{-\frac{{z_j}^2}{2}\Big\}\} dz_1 \ldots dz_b\}}$$

로 주어진다. 그러나 이 적분은 수값으로만 계산이 행해진다. 수치적분(numerical integration)을 행하는 한가지 방법은 Gauss-Hermite 공식을 이용하여 근사적으로 계산된다. 따라서, 이 MLLH를 미지모수들에 대해 편미분하여 얻은 연립방정식들의 해가 최우추정치들이 된다. 이들 방정식은 미지모수들에 대해 비선형이므로 numerical algorithm을 이용하여 구하게 된다.

4. 생성자료의 예

다가자료의 분석을 위하여 혼합효과모형을 이용하는 경우의 예로써 다음과 같은 연구를 가정해 본다. 종합병원에서 근무하는 병리검사자들을 대상으로 판독이 쉽지 않은 한 검사 결과물에 대하여 판독능력을 알아 본다고 하자. 판독결과는 음성, 양성, 그리고 판독불능으로 분류된다. 결과물의 판독능력은 재직연수와 관련이 있다고 생각할 때 재직연수는 설명변수로 간주된다. 재직연수가 검사물의 판독율에 미치는 효과를 추론하기 위하여 종합병원에서 근무하는 병리검사자들을 대상으로 단순집락표집으로 표본을 추출한다고 하자. 즉, 종합병원들의 집단에서 일부병원을 임의로 추출하고 추출된 병원에서 근무하는 병리검사자들을 대상으로 판독력을 검사한다. 이때 판독율에 영향을 미치는 요인으로 병원간의 변동을 생각할 수 있고, 추출된 병원 j의 효과는 $N(0, \sigma_h^2)$의 분포를 따르는 확률효과로 간주된다. 따라서, 판독결과는 세개의 범주를 갖는 명목형 반응변수이고, 재직연수는 네개의 범주를 갖는 순서형 설명변수이며, 종합병원은 명목형 변수이나 관측된 세 병원은 명목형 확률변수의 관측값들이다.단순집락표집법에 의해 추출된 표본자료의 결과표가 다음과 같다고 하자.

		판독결과		
병원	재직연수	음성	양성	판독불가
A	1년미만	1	5	10
	1-3	3	4	8
	3-5	5	2	7
	5년이상	5	2	6
B	1년미만	2	7	9
	1-3년	2	5	9
	3-5	4	3	8
	5년이상	4	1	5
C	1년미만	1	4	11
	1-3	3	4	9
	3-5	4	2	10
	5년이상	5	2	7

표 4.1: 검사물의 생성자료

위의 자료를 분석하기 위한 모형으로 다음 모형을 가정한다.

$$log(\frac{P(Y=h|ij)}{P(Y=|ij)}) = \alpha_h + \sigma_h z_i + \beta_h u_j \quad (4.1)$$
$$i = A, B, C, j = 1, 1-3, 3-5, 5, h = ,.$$

판독불가의 반응범주를 기본범주로 하는 기본범주 로짓모형이다. 위의 모형은 두 유형의 로짓들에 대한 방정식들을 포함하고 있다. 즉,

$$log(\frac{P(Y=)}{P(Y=)}) = logit(\pi_{1|ij}) = \alpha_1 + \sigma_1 z_i + \beta_1 u_j \quad (4.2)$$
$$log(\frac{P(Y=)}{P(Y=)}) = logit(\pi_{2|ij}) = \alpha_2 + \sigma_2 z_i + \beta_2 u_j$$

단, u_i는 재직연수의 범주들에 대한 평균점수 0.5, 2, 4, 그리고 5.5를 나타낸다.검사물의 생성자료를 분석하기 위하여 모형(10)의 가정하에 모형내 미지모수들을 추정하기 위한 과정을 살펴보기 위하여 표본자료에 대한 우도함수를 LH라 두자. 이때, 우도함수는 다음과 같이 구해진다. 우선 요인 A의 i번째 수준, 요인 B의 j번째 수준에서 표본의 크기를 n_{ij}, 음성결과의 관측도수를 y_{0ij} 그리고 양성결과의 관측도수를 y_{1ij}라 두자. 두 요인의 결합수준 (i,j)에서 검사물 결과에에 대한 관측도수들의 확률분포는 모형(10)을 적용할 때 다음의 다항분포를 따르게 된다.

$$\frac{n_{ij}!}{y_{0ij}!y_{1ij}!(n_{ij}-y_{0ij}-y_{1ij})!}(\frac{\exp(\alpha_1+\sigma_1 z_i+\beta_1 u_j)}{1+\exp(\alpha_1+\sigma_1 z_i+\beta_1 u_j)+\exp(\alpha_2+\sigma_2 z_i+\beta_2 u_j)})^{y_{0ij}} \quad (4.3)$$
$$(\frac{\exp(\alpha_2+\sigma_2 z_i+\beta_2 u_j)}{1+\exp(\alpha_1+\sigma_1 z_i+\beta_1 u_j)+\exp(\alpha_2+\sigma_2 z_i+\beta_2 u_j)})^{y_{1ij}}$$
$$(\frac{1}{1+\exp(\alpha_1+\sigma_1 z_i+\beta_1 u_j)+\exp(\alpha_2+\sigma_2 z_i+\beta_2 u_j)})^{n_{ij}-y_{0ij}-y_{1ij}}$$

위의 분포는 병원 i가 주어졌을 때, 관측도수들의 조건부 분포이다. 즉, 확률변수 Z_i가 z_i일 때의 조건부 분포이다. 확률변수 Z_i가 평균이 0이고 분산이 1인 정규분포를 따른다고 가정할 때, Z_i의 분포를 $\phi(z_i)$라 두자. 이 때, 두 요인들의 모든 결합수준에서 관측되는 도수들의 우도함수 LH는 다음과 같다.

$$LH = \prod_{i=1}^{3}\{\phi(z_i)\prod_{j=1}^{4}\{\frac{n_{ij}!}{y_{0ij}!y_{1ij}!(n_{ij}-y_{0ij}-y_{1ij})!} \quad (4.4)$$
$$(\frac{\exp(\alpha_1+\sigma_1 z_i+\beta_1 u_j)}{1+\exp(\alpha_1+\sigma_1 z_i+\beta_1 u_j)+\exp(\alpha_2+\sigma_2 z_i+\beta_2 u_j)})^{y_{0ij}}$$
$$(\frac{\exp(\alpha_2+\sigma_2 z_i+\beta_2 u_j)}{1+\exp(\alpha_1+\sigma_1 z_i+\beta_1 u_j)+\exp(\alpha_2+\sigma_2 z_i+\beta_2 u_j)})^{y_{1ij}}$$
$$(\frac{1}{1+\exp(\alpha_1+\sigma_1 z_i+\beta_1 u_j)+\exp(\alpha_2+\sigma_2 z_i+\beta_2 u_j)})^{n_{ij}-y_{0ij}-y_{1ij}}\}\}$$

단, $i = 1, 2, 3$은 병원의 세 수준을 나타내고 $j = 1, 2, 3, 4$는 재직연수의 네 수준을 나타낸다. 우도함수 LH는 정규확률변수 Z_i들을 포함하고 있다. 따라서, 모수들의 최우추정치를 구하기 위하여 LH를 Z_i들에 대하여 적분하여 주변우도함수를 구한다. 주변우도함수를 MLH라 두면,

$$MLH = \prod_{i=1}^{3}\{\int \phi(z_i)[\prod_{j=1}^{4}\{\frac{n_{ij}!}{y_{0ij}!y_{1ij}!(n_{ij}-y_{0ij}-y_{1ij})!} \tag{4.5}$$
$$(\frac{\exp(\alpha_1+\sigma_1 z_i+\beta_1 u_j)}{1+\exp(\alpha_1+\sigma_1 z_i+\beta_1 u_j)+\exp(\alpha_2+\sigma_2 z_i+\beta_2 u_j)})^{y_{0ij}}$$
$$(\frac{\exp(\alpha_2+\sigma_2 z_i+\beta_2 u_j)}{1+\exp(\alpha_1+\sigma_1 z_i+\beta_1 u_j)+\exp(\alpha_2+\sigma_2 z_i+\beta_2 u_j)})^{y_{1ij}}$$
$$(\frac{1}{1+\exp(\alpha_1+\sigma_1 z_i+\beta_1 u_j)+\exp(\alpha_2+\sigma_2 z_i+\beta_2 u_j)})^{n_{ij}-y_{0ij}-y_{1ij}}\}]dz_i\}$$

로 표현된다. 위 MLH를 구하기 위한 적분은 수값으로만 계산되므로 수치적분을 위해 Gauss-Hermite 공식을 이용한다. 이 공식은 적절히 선택된 m개의 구적점(quadrature point)과 관련된 가중값을 필요로 한다. 수치적분에 의해 근사적으로 구해진 주변우도함수를 AMLH라 둘 때 AMLH는 다음과 같다.

$$AMLH = \pi^{-\frac{3}{2}}\prod_{i=1}^{3}\{\sum_{k=1}^{m} w_k[\prod_{j=1}^{4}\{\frac{n_{ij}!}{y_{0ij}!y_{1ij}!(n_{ij}-y_{0ij}-y_{1ij})!} \tag{4.6}$$
$$(\frac{\exp(\alpha_1+\sigma_1 x_k\sqrt{2}+\beta_1 u_j)}{1+\exp(\alpha_1+\sigma_1 x_k\sqrt{2}+\beta_1 u_j)+\exp(\alpha_2+\sigma_2 x_k\sqrt{2}+\beta_2 u_j)})^{y_{0ij}}$$
$$(\frac{\exp(\alpha_2+\sigma_2 x_k\sqrt{2}+\beta_2 u_j)}{1+\exp(\alpha_1+\sigma_1 z_i+\beta_1 u_j)+\exp(\alpha_2+\sigma_2 x_k\sqrt{2}+\beta_2 u_j)})^{y_{1ij}}$$
$$(\frac{1}{1+\exp(\alpha_1+\sigma_1 x_k\sqrt{2}+\beta_1 u_j)+\exp(\alpha_2+\sigma_2 x_k\sqrt{2}+\beta_2 u_j)})^{n_{ij}-y_{0ij}-y_{1ij}}\}]\}$$

여기서 x_k는 구적점을 나타내고 w_k는 x_k와 관련된 가중값을 나타내고 있다. AMLH를 대수변환한 후 미지모수들에 대해 편미분하여 구한 방정식들은 미지모수들에 관하여 비선형이기 때문에 최우추정치들을 얻기 위한 numerical algorithm으로 Nelder and Mead(1965)의 심플렉스 방법을 이용한다. 확률요인 A와 고정요인 B의 두 요인을 포함하고 있는 다원분할표에서 반응변수의 관측값이 명목형의 다가범주로 주어질 때의 자료분석예로써 검사물의 생성자료를 이용하였다. 자료분석을 위한 모형(10)의 가정으로 부터 구한 모형내 모수들의 추정값과 해당하는 표준오차들은 다음과 같다.$\hat{\alpha_1}$=-2.467(0.079), $\hat{\beta_1}$=0.286(0.021),$\hat{\alpha_2}$=-1.001(0.053), $\hat{\beta_2}$=-0.190(0.017),$\hat{\sigma_1}$= 0.000007(0.022), $\hat{\sigma_2}$=0.000020(0.028)

자료분석을 위한 연속모형(10)의 적합성을 알아보기 위한 측도로써 이용되는 이탈도의 값은 71.27이고 해당하는 자유도는 18이 된다. 평균이탈도가 1에서 상당히 떨어져 있으므로 위 생성자료에 모형(10)의 혼합효과 모형이 적합한 모형이라고 판단되지는 않으나 다원분할표에서 고정효과와 확률효과를 추정하는 방법의 한 예로써 제시하고 있다.

5. 결론

본 논문은 실험 또는 조사를 통하여 얻게되는 도수들의 다원분할표에서 개체에 대한 반응변수가 다가의 범주로 관측되고 개체의 반응에 영향을 미치는 독립변수들중 일부는 고정요인들이고 일부는 확률요인들로 분류된 경우의 범주형 자료를 분석하기 위한 모형을 제시하고 있다. 제시된 모형내 미지모수들을 추정하기 위한 방법으로 주변우도함수를 이용한

최우추정법으로 구체적인 예를 통하여 모수들의 최우추정치와 이들의 표준오차를 구하는 방법을 제공하고 있다. 그러나 본 논문의 주안점은 개체의 반응변수가 셋 이상의 다가범주로 주어지는 다가의 명목형 변수이고, 반응에 영향을 미치는 독립변수들이 고정요인과 확률요인들이며 고정요인은 양적수준을 갖는 순서형 변수로 간주할 수 있을 때, 모형설정을 하는 데 두고 있다. 반응변수의 관측범주들이 단순한 분류를 나타내는 명목형의 범주이므로 기준범주(baseline category)를 이용한 일반화로짓(generalized logit)을 이용하여 모형을 설정할 수 있는 점에 착안하여 모형을 제시하게 되었다.

참고문헌

[1] Agresti, Alan. (1990). *Categorical data analysis*, John Wiley and Sons, Inc., New York.

[2] Cox, D. R. and Snell, E. J. (1989). *Analysis of binary data* (2nd edition), Chapman and Hall, London.

[3] Im, S. and Gianola, D. (1988). Mixed models for binomial data with an application to lamb mortality. *Applied Statistics*, Vol. **37**, 196-204.

[4] McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models* (2nd edition), Chapman and Hall, London.

[5] Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, Vol. **7**, 308-313.

[6] Williams, D. A. (1982a). Extra-binomial variation in logistic linear models. *Applied Statistics*, Vol. **31**, 144-148.

[2001년 9월 접수, 2002년 3월 채택]

A generalized logit model with mixed effects for categorical data*

Jaesung Choi[1)]

ABSTRACT

This paper suggests a generalized logit model with mixed effects for analysing frequency data in multi-contingency table. In this model nominal response variable is assumed to be polychotomous. When some factors are fixed but considered as ordinal and others are random, this paper shows how to use baseline-category logits to incoporate the mixed-effects of those factors into the model. A numerical algorithm was used to estimate model parameters by using marginal log-likelihood.

Keywords: polychotomous data, baseline-category logits,mixed-effects.

* The present research has been conducted by the Bisa Research Grant of Keimyung University in 2000.

1) Professor, Department of Statistics, Keimyung University, 1000 Sindang-Dong, Dalseogu, Taegu 704-701, Korea.

1) 본 연구는 2000년도 계명대학교 비사연구기금으로 이루어졌음

1) (704-701) 대구광역시 달서구 신당동 1000, 계명대학교 통계학과 교수