Journal of Korean
Data & Information Science Society
2004, Vol. 15, No. 2, pp. 339～345

A Mixed Model for Oredered Response Categories

Jaesung Choi[1)]

Abstract

This paper deals with a mixed logit model for ordered polytomous data. There are two types of factors affecting the response varable in this paper. One is a fixed factor with finite quantitative levels and the other is a random factor coming from an experimental structure such as a randomized complete block design. It is discussed how to set up the model for analyzing ordered polytomous data and illustrated how to estimate the paramers in the given model.

Keywords : mixed model, ordered data, polytomous data, random factor.

1. 서론

개체에 대한 반응이 순서형 범주의 다가자료로 주어지고 반응에 영향을 미치는 두 가지 유형의 요인, 즉, 고정요인 과 확률요인이 존재할 때 이들이 자료에 미치는 영향을 조사하게 된다. 실험 또는 관측조사로부터 주어지는 자료들은 다양한 요인들을 포함하게 되고 이들 요인들을 모형에 포함시켜 그 효과를 추론하는 것이 때로는 간단하지가 않게 된다. 특히 반응에 영향을 미치는 확률요인이 실험계획으로부터 주어질 때 좀더 복잡한 경우의 자료분석 모형이 필요하게 된다.

본 논문에서는 실험계획과 관련된 확률요인과 처치로서의 고정요인을 고려한 혼합모형에서의 자료분석방법을 논의하고자 한다. 또한 순서형 자료를 고려하고 있기 때문에 수집된 자료는 적어도 셋이상의 유한개의 범주로 구성되는 다가자료를 의미한다. 다가자료(polytomous data)는 자료의 특성상 이가자료(binary data) 또는 이항자료(grouped binary data)와는 달리 자료구조의 복잡성 때문에 분석방법도 용이하지 않음을 알 수 있다. 순서형 다가자료의 구조적 특성을 고려한 다양한 모형들 및 분석방법들은 McCullagh and Nelder(1989) 와 Agresti(1990)에서 논의되고 있다. Im and Gianola(1988)는 이원지분계획으로부터 발생하는 분산성분들을 추정하기 위하여 이항자료에 대한 혼합모형을 다루고 있으나 순서형 다가자료를 분석하기 위한 혼합모형에

---

1) Professor, Department of Statistics, Keimyung University. Taegu, 704-701, Korea
E-mail: jschoi@kmu.ac.kr.

관한 논의는 찾아보기가 쉽지 않다. 개체 또는 실험단위의 반응에 대한 단순척도(pure scale)의 관측반응이 다가의 범주로 주어질 때, 실험 또는 조사로부터 수집된 자료는 다가자료를 구성하게 되고 이들이 반응범주의 도수로 표현되면 다항자료(multinomial data)라 한다.

본 연구는 관심모집단의 개체에 대한 관측이 순서형 다가범주중 하나로 관측되고 고정요인들의 효과를 추론하기 위한 실험계획에서 개체의 반응에 영향을 미치는 확률요인을 가정할 때 혼합모형의 제시와 함께 모수내 미지모수들을 추론하는 방법을 논의한다.

2. 모형

순서형 반응변수의 관측범주에 대한 확률을 분석하기 위한 모형설정을 위하여 필요한 몇가지 조건들을 가정해 본다. 먼저, 관심모집단내 개체에 대한 반응이 순서형의 다범주(multi-category)로 관측되고 각 반응범주의 확률에 영향을 미치는 독립변수들로 두 요인 A 와 B를 고려한다. 요인 A는 $i=1,2,\ldots,a$ 개의 수준들로 이루어진 고정요인(fixed factor)이고 요인 B는 실험에 이용되는 실험단위들의 이질성으로 인해 실험단위들을 $j=1,2,\ldots,b$ 개의 동질적인 그룹으로 분류하는 블록(block)요인이다. 각 블록이 a개의 동질적인 실험단위들을 갖는다고 가정할 때 이용되는 실험계획은 완전임의 블록설계(randomized complete block design)이다. 그리고 개체에 대한 반응은 $k=1,2,\ldots,l$ 개의 순서형 범주들로 주어지는 반응변수 Y로 나타낸다. 연구자의 관심모집단에서 개체의 세가지 특성 A, B와 Y에 대한 조사는 세변수 A, B와 Y의 결합확률분포로 표현되거나 요인 A와 요인 B의 주어진 수준하에서 조건부 확률분포로도 표현될 수 있다. 즉, { $\pi_{k|ij}$}이다. 순서형 반응변수 Y의 각 범주에 속할 확률에 영향을 미치는 두 요인들의 효과를 추론하기 위하여 실험을 행한후 자료를 수집한다. 자료수집을 위해 n_{ij}를 블럭요인 B의 블록 j에서 요인 A의 처치 또는 수준 i 가 행해진 실험단위들의 수라 두자. 따라서 n_{ij}개 실험단위에서 순서형 반응변수 Y의 l 개 범주들의 관측도수는 n_{ijk}로 주어지고 { n_{ijk} }는 다항분포를 따르게 된다. 이들 순서형 자료를 분석하기 위한 모형은 고정요인을 측정척도에 따른 범주형 변수로 분류할 때 몇가지 가능한 모형들을 기술할 수 있다. 첫 째는 고정요인이 명목형 변수일 때 자료모형은 다음과 같이 기술된다.

$$g(P(Y=k|ij))=\alpha_k+\ \beta_{ik}{}^A+\ \beta_{jk}{}^B \tag{2.1}$$
$$i=1,2,\ldots,a,\ \ j=1,2,\ldots b,\ \ k=1,2,\ldots,l-1.$$

여기서 $g(\cdot)$는 연결함수이고, α_k는 반응변수 Y가 범주 k로 반응할 때의 절편을 나타내며 { $\beta_{ik}{}^A$ }는 요인 A의 고정효과를 { $\beta_{jk}{}^B$ }는 블록요인 B의 확률효과를 나타낸다.

{ β_{jk}^{B} }는 확률효과들이므로 $N(0,\ \sigma_B^2)$을 따른다고 가정한다. 여기서 유의할 점은 요인 B의 b개 수준들은 블록집단에서 임의로 추출된 블록들로 가정되므로 동일분산을 가정하고 있다. 두번 째는 요인 A가 순서형 변수일 때의 모형이다.

$$g(P(Y=k|ij)) = \alpha_k + \lambda_k u_i + \beta_{jk}^{B} \tag{2.2}$$
$$i=1,2,\ldots,a,\ j=1,2,\ldots b,\ k=1,2,\ldots,l-1.$$

단, { u_i }는 요인 A의 수준들과 동일한 순서를 갖는 단조점수들이고 { λ_k }는 연결함수 g로 변환된 값들에 대하여 각 범주에 따른 기울기이다. 반응변수가 셋이상의 다범주를 갖는 순서형 변수이므로 다양한 변환함수를 이용할 수 있다. Agresti(1990)는 반응범주들이 자연스러운 순서를 가질 때, 그 순서를 이용할 수 있는 세 가지 유형의 로짓변환을 소개하고 있다. 그 세가지는 인접범주 로짓(adjacent-categories logits), 연속비 로짓(continuation-ratio logits) 그리고 누적로짓(cumulative logits)이다. 본 논문은 인접범주 로짓을 이용한 혼합효과 모형을 자료에 적합시켜 보고자 한다. 인접범주 로짓은 다음과 같이 정의한다.

$$L_k = \log \frac{\pi_k(x)}{\pi_{k+1}(x)}\ ,\ k=1,2,\ \ldots\ l-1.$$

인접범주 로짓을 이용할 때 식(2.1)은

$$L_{k|ij} = \alpha_k + \beta_{ik}^{A} + \beta_{jk}^{B},\ k=1,2,\ \ldots\ ,l-1. \tag{2.3}$$

으로 주어진다. 이때, 서로 다른 요인들의 수준에서 동일 로짓의 차는

$$\begin{aligned} L_{k|ij} - L_{k|i'j'} &= \alpha_k + \beta_{ik}^{A} + \beta_{jk}^{B} - \alpha_k - \beta_{i'k}^{A} - \beta_{j'k}^{B} \\ &= (\beta_{ik}^{A} - \beta_{i'k}^{A}) + (\beta_{jk}^{B} - \beta_{j'k}^{B}) \end{aligned}$$

이고 $E(L_{k|ij} - L_{k|i'j'}) = \beta_{ik}^{A} - \beta_{i'k}^{A}$ 임을 보여주고 있다. 단, $i \neq i'$, $j \neq j'$ 이다.
인접범주 로짓을 이용한 식(2.2)는

$$L_{k|ij} = \alpha_k + \lambda_k u_i + \beta_{jk}^{B},\ k=1,2,\ \ldots\ ,l-1. \tag{2.4}$$

이고 동일 블록내 고정요인 A의 서로 다른 두 수준간의 차는

$$
\begin{aligned}
L_{k|ij} - L_{k|i'j} &= \alpha_k + \lambda_k u_i + \beta_{jk}{}^{B} - \alpha_k - \lambda_k u_{i'} - \beta_{jk}{}^{B} \\
&= \lambda_k (u_i - u_{i'})
\end{aligned}
$$

임을 보여준다. 단, $i \neq i'$, $i = 1, 2, \ldots, a$. 이다.
두 인접범주 로짓의 차이가 로그승산비임을 감안할 때, 로그승산비는 단순히 고정요인 A의 두 수준간에 효과차를 나타내고 있다. 또한 식(2.4)는 모수 λ_k가 양수일 때, 각 로짓은 u_i가 증가함에 따라 커지게 되고 따라서 각 범주확률이 증가하게 된다.

3. 자료구조

2절 모형의 논의에서 고려된 실험모집단에서 추출된 실험단위(experimental unit)를 이용하여 실험한 후 주어지는 자료구조를 생각해 보기로 한다. 실험에 이용될 실험단위들의 이 질성으로 인하여 동질적인 실험단위들을 블록화하고 각 블록내에서 실험단위들이 비교하고자 하는 한 고정요인의 모든 수준에 임의로 배정되는 것을 가정하였기 때문에 이용된 실험계획은 확률화 완비계획(randomized complete block design)이다. 비교하고자 하는 모든 처치를 포함하고 있는 각 블록은 임의성에 다른 실험단위들의 반응에 확률효과를 미치는 것으로 가정한다. 이러한 가정하에 자료구조는 다음 표로 요약된다.

<표 3.1> 확률화 완비계획하의 자료구조를 나타내는 다가자료표

		반응범주						
블록(B)	고정요인(A)	1	2	$\cdots$	k	$\cdots$	l
1	$\cdots$				$\cdots$		
$\cdots$	$\cdots$						
j	i	y_{ij1_1}	y_{ij2_1}	$\cdots$	y_{ijk_1}	$\cdots$	y_{ijl_1}
		y_{ij1_2}	y_{ij2_2}	$\cdots$	y_{ijk_2}	$\cdots$	y_{ijl_2}
					$\cdots$		
		y_{ij1_t}	y_{ij2_t}	$\cdots$	y_{ijk_t}	$\cdots$	y_{ijl_t}
$\cdots$	$\cdots$				$\cdots$		
b	$\cdots$				$\cdots$		

<표 3.1>에서 관측값 y_{ijk_t}는 j번째 블록에서 처치 또는 고정요인 A의 수준 i가

행해진 t번째 실험단위의 관측반응이 범주 k임을 나타내고 있다. 또한 표의 자료로부터 각 블록내 실험단위의 수는 ac 개로 동일 수의 실험단위들을 포함시키고 있음을 알 수 있다. 단, c는 양의 상수이다. 여기서 y_{ijk_t}는 다가반응변수의 l 개 범주중 하나로 관측되기 때문에 이들 관측값들은 다가자료를 구성하게 된다. y_{ijk}를 j 번째 블록내 처치 i가 행해진 것중 범주 k인 실험단위들의 수라 두자. 따라서, j 번째 블록내 처치 i가 행해진 실험단위들의 수를 n_{ij}라 두면, 확률벡타 $Y_{ij}=(Y_{ij1}, Y_{ij2}, \ldots, Y_{ijl})'$은 시행회수 n_{ij}와 $\pi_{ij}=(\pi_{ij1}, \pi_{ij2}, \ldots, \pi_{ijl})'$인 다항분포를 따르게 된다. 단, π_{ijk}는 개체의 반응변수 Y_{ijk_t}가 범주 k로 관측될 확률이다.

4. 자료의 예

순서형 다가자료의 분석을 위하여 혼합효과 모형을 이용하는 경우의 예로써 다음과 같은 연구를 가정해 본다. 한 임업시험가가 관심묘목의 생장률에 영향을 미치는 것으로 간주되는 요인 A의 다섯 수준에서 그 효과를 추론하는 데 관심을 갖고 있다 하자. 묘목시험장으로 전국에서 세 시험장을 임의로 선정한다. 각 시험장에서는 40그루씩 임의로 다섯 수준에 배정한다. 일정기간 뒤 개별묘목에 대한 관측값은 생장도를 나타내는 기준에 근거하여 생장도가 보통, 좋음, 우수함의 세가지 범주로 관측된다. 이 예는 전절에서 논의된 자료구조를 취하고 있다. 왜냐하면, 묘목의 생장도는 실험단위 또는 개체의 반응변수로 세 가지 범주간에 순서가 고려된 순서형 반응변수임을 알 수 있다. 관심묘목의 생장도에 대한 관측범주간에 순서가 주어졌기 때문에 인접범주 확률을 이용한 로짓변환으로 처치효과를 추론해 볼 수 있다. 생성된 자료를 이용하여 구체적으로 누적로짓 혼합모형을 적합시켜 모형의 타당성과 추론 방법을 살펴보기로 한다. 관심묘목의 생장률을 분석하기 위한 생성자료가 <표 4.1>에 주어진다.

<표 4.1> 묘목의 생장률에 대한 생성자료

		생장도		
시험장	요인A	보통	좋음	우수
1	2	20	15	5
	5	20	12	8
	7	15	15	10
	9	13	12	15
	12	8	14	18
2	2	20	13	7
	5	18	14	8
	7	13	16	11
	9	10	15	15
	12	8	7	25
3	2	22	13	5
	5	21	11	8
	7	17	14	9
	9	11	8	21
	12	6	9	25

위 자료를 분석하기 위하여 식(2.4)를 적합시켜 보기로 한다. 식(2.4)는 다음과 같이 변형될 수 있다.

$$L_{k|ij} = \alpha_k + \lambda_k u_i + \sigma_B z_j \tag{4.1}$$

단, $k = 1, 2$, 이고 z_j는 $N(0,1)$인 표준정규변수이다. $j = 1, 2, 3$ 이다.

블록과 처리가 주어졌을 때, 관측도수들의 분포는 다항분포를 따르게 된다. 따라서 블록 j, 처리 i에서 각 범주내 관측도수를 n_{ijk}라 두면 두 요인의 모든 수준결합에서 관측도수들의 분포는 곱다항분포를 따르게 된다.

$$\prod_{j=1}^{3} \prod_{i=1}^{5} \left\{ \frac{n_{ij}!}{n_{ij1}!\, n_{ij2}!\, n_{ij3}!} \; \pi_{1|ij}{}^{n_{ij1}} \; \pi_{2|ij}{}^{n_{ij2}} \; \pi_{3|ij}{}^{n_{ij3}} \right\} \tag{4.2}$$

위 식(4.2)에 혼합모형식(4.1)을 대입하면 아래의 우도함수를 얻게된다.

$$\prod_{j=1}^{3} \prod_{i=1}^{5} \left\{ \frac{n_{ij}!}{n_{ij1}!\, n_{ij2}!\, n_{ij3}!} \left\{ \frac{\exp(\alpha_1 + \lambda_1 u_i + \alpha_B z_j)}{1 + \exp(\alpha_1 + \lambda_1 u_i + \alpha_B z_j) + \exp(\alpha_2 + \lambda_2 u_i + \alpha_B z_j)} \right\}^{n_{ij1}} \right.$$
$$\left\{ \frac{\exp(\alpha_2 + \lambda_2 u_i + \alpha_B z_j)}{(1 + \exp(\alpha_1 + \lambda_1 u_i + \alpha_B z_j) + \exp(\alpha_2 + \lambda u_2 + \alpha_B z_j))} \right\}^{n_{ij2}}$$
$$\left. \left\{ \frac{1}{1 + \exp(\alpha_1 + \lambda_1 u_i + \sigma_B z_j) + \exp(\alpha_2 + \lambda_2 u_i + \sigma_B z_j)} \right\}^{n_{ij3}} \right\}$$

우도함수를 이용하여 모수들의 최우추정값들을 구하기 위하여 Gauss-Hermite 공식을 이용하여 주변우도함수를 구한다. 그 다음 주변우도함수를 대수변환한 후 미지모수들에 대해 편미분하여 얻은 연립방정식들의 해는 Nelder and Mead(1965)의 심플렉스 방법을 이용해 구해진다. 구해진 해는 다음과 같다.

$\widehat{\alpha_1}$=2.048(0.0141), $\widehat{\alpha_2}$=1.403(0.0087), $\widehat{\lambda_1}$=-0.260(0.0004), $\widehat{\lambda_2}$=-0.184(0.0003) 이고 $\widehat{\sigma_B}$=0.00002(0.0165) 이다.

괄호안은 추정량의 분산에 대한 추정값을 나타내고 있다. 인접범주 로짓혼합모형(4.1)의 적합성을 알아보기 위한 측도로써 이용되는 이탈도의 값은 127.8이고 해당하는 자유도는 25이다. 평균이탈도가 1로부터 상당히 떨어져 있으므로 여러 가지 다양한 모형을 적합시켜 자료에 적합한 모형을 살펴볼 수는 있겠으나 순서형 다가자료에 대한 혼합모형을 적합시키는 방법의 한 예로써 제시하고 있다.

5. 결론

본 논문은 실험 또는 관측조사를 통하여 수집되는 자료가 다가의 순서형 자료이고, 개체의 반응에 영향을 미치는 요인들이 고정요인과 확률요인 둘다 포함하고 있는 경우를 가정하고 있다. 여기서 고정요인은 유한개의 수준으로 구성된 양적변수이고 확률요인은 실험단위들의 이질성으로 인한 실험계획으로부터 발생하는 요인을 고려하고 있다. 다가의 순서형 반응범주들에 대한 변환은 인접범주 로짓을 이용한 로짓모형을 제시하고 모형내 미지모수들의 추정값과 표준오차를 구하는 방법을 논의하였다.

참고문헌

1. Abramowitz, M. and Stegun, I. (1972). Handbook of mathematical functions, p.924, Dover Publications, New York.
2. Agresti, Alan. (1990). Categorical data analysis, John Wiley and Sons, Inc., New York.
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.

[2004년 1월 접수, 2004년 4월 채택]