Journal of Korean
Data & Information Science Society
2006, Vol. 17, No. 1, pp. 123 ~ 130

A Cumulative Logit Mixed Model for Ordered Response Data

Jaesung Choi[1)]

Abstract

This paper discusses about how to build up a mixed-effects model using cumulative logits when some factors are fixed and others are random. Location effects are considered as random effects by choosing them randomly from a population of locations. Estimation procedure for the unknown parameters in a suggested model is also discussed by an illustrated example.

Keywords : Cumulative logit, Mixed model, Ordered data, Random effects

1. 서론

관측조사 또는 실험연구로부터 개체 또는 실험단위의 반응이 유한개의 관측값들로 주어지는 범주형 반응들일 때 이들 반응에 영향을 미치는 여러 가능한 독립변수들을 생각할 수 있다. 개체에 대한 반응은 측정척도로 분류할 때 명목형, 순서형 그리고 구간형으로 분류된다. 측정척도에 따른 관측반응의 유형에 따라 자료분석 방법은 달라진다. 따라서, 본 논문은 개체의 반응이 측정척도에 따라 순서형으로 분류되고 관측반응에 영향을 미치는 다수의 독립변수들이 있음을 전제하고 있다. 실험단위 또는 개체의 반응변수가 순서형의 반응변수로 간주되는 다가의 범주형 변수일 때, 단일 반응변수의 관측자료를 분석하기 위한 로짓 변환으로는 인접범주 로짓(adjacent-categories logits), 연속비 로짓(continuation-ratio logits) 그리고 누적로짓(cumulative logits)등이 있다. 개체의 반응에 영향을 미칠 수 있는 독립변수들은 유한개의 범주를 갖는 질적변수들인 요인들과 연속적인 값을 갖는 것으로 간주되는 양적변수들인 공변량들이 있다. 질적변수로 취급되는 요인들의 유형에는 두 가지가 있다. 하나는 고정요인이고 다른 하나는 확률요인이다. 고정요인들은 대개 반응에 영향을 미치는 처치들이고 이들은 고정효과를 갖는 것으로 간주된다. 실험 또는 관측연구에

1) 대구광역시 달서구 신당동 1000번지 계명대학교 통계학과 교수
E-mail : jschoi@kmu.ac.kr

서 확률요인들로 간주되는 변수들의 수준들은 수준들의 모집단에서 확률표본으로 추출된 일부 수준들임을 의미하고 있다. 따라서 표본으로 다른 수준들이 추출되면 개체의 반응에 영향을 미치는 효과는 달라지게 된다. 이는 수준들의 효과가 어떤 미지의 고정효과가 아닌 수준마다 차이가 있을 수 있는 변동효과를 말하고 있다. 따라서, 수준들의 집단에서 개별수준의 효과들이 관측되면 하나의 분포를 가정할 수 있게 되고 이들 분포는 일반적으로 평균이 0이고 상수분산을 갖는 정규분포를 따르는 것으로 가정한다. 이때, 임의로 주어지는 개별수준의 효과를 확률효과라 부른다. 확률요인들의 발생은 관측연구에서는 관측단위들을 얻기 위한 표본설계에서 생길 수 있고 실험연구에서는 자료수집을 위한 실험설계로부터 발생할 수 있다. 따라서, 본 연구는 관측연구 또는 실험연구로부터 수집되는 자료가 범주형 자료이고, 개체에 대한 반응이 순서형 범주의 다가자료로 주어지며 반응에 영향을 미치는 두 가지 유형의 요인, 즉, 고정요인 과 확률요인이 존재할 때 이들이 자료에 미치는 분석모형의 설정에 관심을 두고 있다. 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 표본추출계획으로부터 주어질 때 좀더 복잡한 경우의 자료분석 모형이 필요하게 된다. 모형에 근거한 자료분석 방법은 모형에 근거하지 않은 자료분석방법 보다 체계적이고 효과적인 자료분석 방법을 제공하게 된다. 따라서 자료분석을 위한 모형은 반응에 영향을 미치는 요인들의 유형에 따라 세 가지로 분류된다. 모형이 고정요인만을 고려하는 있는 경우의 고정효과모형(fixed-effects models), 확률요인만을 고려하고 있는 경우의 확률효과모형(random-effects models)과 두 유형의 요인들을 모두 고려한 경우의 혼합효과모형(mixed-effects models)이다.

본 논문에서는 표본추출계획 또는 실험계획과 관련된 확률요인과 처치로서의 고정요인을 고려한 혼합효과 모형에서의 자료분석 방법을 논의하고자 한다. 또한 순서형 자료를 고려하고 있기 때문에 수집된 자료는 적어도 셋 이상의 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 누적로짓 혼합모형에 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 연구에서 표본추출단계 또는 실험계획단계로 부터 개체들의 반응에 영향을 미치는 확률요인들이 발생하게 될 때 혼합모형의 제시와 함께 모형내 미지모수들을 추론하는 방법을 논의한다.

2. 모형의 가정

개체의 반응과 관련한 몇 가지 가정들을 생각해 본다. 첫째 처치가 행해진 개체의 반응이 유한개의 순서가 주어진 범주를 갖는 관측값으로 나타난다 가정하자. 둘째로 순서형 변수의 각 반응범주에 속할 확률이 관심요인들에 의해 어떻게 영향을 받는 가를 파악하기 위한 변환으로 누적로짓 변환을 가정한다. 셋째로 확률요인들은 실험의 설계구조나 개체들의 표본추출계획으로부터 발생하는 요인들임을 가정한다. 확률요인은 개체의 반응에 영향을 미치는 독립변수로 유한개의 수준만이 실험에 고려되고 이 유한개의 수준들이 모집단으로부터 임의로 추출되었다고 가정하기 때문에 그 효과들은 확률효과들로 간주된다. 이는 확률요인의 수준들이 반응을 나타내는 실험단위 또는 개체들에 행해졌음을 의미한다. 이러한 확률요인은 실험단위들의 표본추출방법으로부터 발생할 수 있다. 이러한 가정과 관련한 혼합모형의 제시를 위하여 다음과 같은 실험환경을 가정한다. 여기서 혼합모형의 의미는 처치구조에서 고정효과를 갖는 일부 고정요인들이 있고 하나 이상의 분산성분을 갖는 모형을 의미한다. 실험환경의 가정으로 개체 또는 실험단위의 반응이 순서형의 다범주(multi-category)로 관측되고 각 반응범주의 확률에 영향을 미치는 독립변수로 두개의 요인 A와 B를 고려한다. 요인 A는 $i=1,2,\ldots,a$ 개의 수준들로 이루어진 고정요인(fixed factor)이고 요인 B는 표본추출방법과 관련하여 발생하는 확률요인으로 가정한다. 요인 B는 $j=1,2,\ldots,b$개의 수준들로 이루어진다. 개체에 대한 반응은 $k=1,2,\ldots,l$ 개의 순서형 범주들로 주어지는 반응변수 Y로 나타낸다.

연구자의 관심모집단에서 개체의 세 가지 특성 A, B와 Y에 대한 조사는 세 변수 A, B 와 Y의 결합확률분포로 표현되거나 요인 A와 요인 B의 주어진 수준하에서 조건부 확률분포로도 표현될 수 있다. 즉, $\{\pi_{\text{깨}ij}\}$이다. 순서형 반응변수 Y의 각 범주에 속할 확률에 영향을 미치는 두 요인들의 효과를 추론하기 위하여 실험을 행한 후 자료를 수집한다. 관심모집단에서 자료수집을 위해 지역들의 집단에서 일정크기의 지역 b개를 임의로 선정하여 개체들을 추출한다 하자. 실험에 이용될 개체 또는 실험단위들을 추출하기 위한 이러한 표본추출계획으로 인하여 개체의 반응확률에 영향을 미치는 확률요인이 발생하게 된다. 즉 반응확률에 있어서 지역간의 변동을 나타내는 지역요인이다. 지역요인의 수준이 임의로 추출되기 때문에 확률요인으로 간주된다. 따라서, 개체의 반응에 영향을 미치는 요인들의 효과는 고정요인의 수준이 반응에 영향을 미치는 고정효과와 한 확률요인의 수준들이 반응에 영향을 미치는 확률효과들을 생각할 수 있다. 순서형 반응변수 Y의 관심범주들에 속할 확률들은 이들 요인들의 효과의 정도를 파악하기 위한 혼합모형으로 주어진다. 모형제시를 위한 다원분류표에서 n_{ij}를 지역 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. 자료분석 모형들의 비교를 위해 반응변수가 명목형이고 고정요인이 또한 명목형 변수일 때를 생각해 보기로 한다. 이때의 자료분석모형은 다음과 같이 기술된다.

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

여기서 $g(\cdot)$는 로짓연결함수이고, α_k는 반응변수 Y가 범주 k로 반응할 때의 절편을 나타내며 $\{\beta_{ik}^{A}\}$는 요인 A의 고정효과를 $\{\beta_{jk}^{B}\}$와 $\{\beta_{ijk}^{AB}\}$는 각기 요인 B 와 교호작용 AB 의 확률효과를 나타낸다.

이들 확률효과들은 각기 $N(0,\ \sigma_B^2)$와 $N(0,\ \sigma_{AB}^2)$을 따른다고 가정한다. 다음으로 순서형 자료를 분석하기 위한 모형에서 고정요인이 순서형이면 모형 (2.1)은 다음과 같이 단순화된다.

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

단, $\{u_i\}$는 요인 A의 수준들과 동일한 순서를 갖는 단조점수들이고 λ 는 연결함수 g로 변환된 값들에 대하여 u_i 에 따른 기울기이다. 모형 (2.2)로의 단순화는 고정요인 A의 수준들 간에 순서를 고려함으로써 부여한 상수 u_i 와 확률효과간의 교호작용은 존재하지 않기 때문에 가능하게 된다. 개체에 대한 반응변수로 셋 이상의 다범주를 갖는 순서형 변수를 가정하고 있기 때문에 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 세 가지는 인접범주 로짓, 연속비 로짓 그리고 누적로짓이다. 본 논문은 누적로짓을 이용한 혼합효과 모형을 자료에 적합시켜 보고자 한다. 누적로짓은 다음과 같이 정의한다.

$$L_k = \log ik[F_k(\boldsymbol{x})] = \log \frac{F_k(\boldsymbol{x})}{1-F_k(\boldsymbol{x})},\quad k=1,2,\ \ldots\ l-1.$$

단, $F_k(\boldsymbol{x}) = \pi_1(\boldsymbol{x}) + \ \ldots\ + \pi_k(\boldsymbol{x})$인 범주 k까지의 누적확률을 나타낸다.

누적로짓을 이용할 때 식 (2.2)는

$$L_{k|ij} = \alpha_k + \lambda u_i + \beta_{jk}^{B},\ k=1,2,\ \ldots\ ,l-1. \quad (2.3)$$

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

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

이고 $E(L_{k|ij}-L_{k|i'j'})=\lambda(u_i-u_{i'})$ 임을 보여주고 있다. 단, $i\neq i'$, $j\neq j'$ 이다.

두 누적로짓의 차이가 로그누적승산비임을 감안할 때, 로그누적승산비는 단순히 고정요인 A의 두 수준간에 효과차를 나타내고 있다. 또한 식 (2.3)은 모수 λ가 양수일 때, 각 누적로짓은 u_i가 증가함에 따라 커지게 되고 따라서 각 누적확률이 증가하게 된다. 이것은 반응변수 Y의 낮은 값에 상대적으로 더 많은 확률이 주어짐을 의미한다. 즉, u_i가 클 때 반응변수는 작게 되는 경향을 나타내므로 양수인 모수 λ가 u_i가 커짐에 따라 반응변수 Y의 큰 값이 대응하는 좀 더 일반적인 의미를 갖도록 하기 위해 λ 대신 $-\lambda$ 로 대체한다. 이때, 식 (2.3)은

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

으로 주어진다.

3. 자료의 예

다음은 묘목의 생장률에 대한 자료 최재성(2004)이다. 이 자료는 관심묘목의 생장률에 영향을 미치는 것으로 간주되는 고정요인 A의 다섯 수준에서 그 효과를 추론하는데 관심을 갖고 있다. 묘목시험장으로 전국에서 세 시험장을 임의로 선정한다. 이때 이용된 시험장 추출방법은 표본으로 추출된 지역의 지역효과가 생장률에 영향을 미칠 수 있는 확률효과를 고려할 때 확률요인으로 간주된다. 각 시험장에서는 40그루씩 임의로 세 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 누적확률을 이용한 누적로짓 변환으로 처치효과를 추론해 볼 수 있다. 주어진 자료를 이용하여 구체적으로 누적로짓 혼합모형을 적합시켜 모형의 타당성과 추론 방법을 살펴보기로 한다. 관심묘목의 생장률을 분석하기 위한 생성자료가 <표 3.1>에 주어진다.

<표 3.1> 묘목의 생장률에 대한 생성자료

시험장	요인A	생장도		
		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.

$$L_{k|ij} = \alpha_k - \lambda_k u_i - \sigma_B z_j \tag{3.1}$$

단, $k=1,2$ 이고 z_j는 $N(0,1)$인 표준정규변수이다. $j=1,2,3$ 이다. 생성자료표의 자료를 분석하기 위한 모형으로 식 (3.1)을 이용하기로 한다. 위 자료에서 반응범주는 셋이므로 모형에 이용되는 누적로짓은 두 종류이고 각각은 다음과 같다.

$$\begin{aligned} L_{1|ij} &= \alpha_1 - \lambda_1 u_i - \sigma_B z_j \\ L_{2|ij} &= \alpha_2 - \lambda_2 u_i - \sigma_B z_j \end{aligned} \tag{3.2}$$

시험장 j 그리고 요인 A의 수준 i에서 각 범주내 관측도수를 n_{ijk} 라 두면 세 요인의 모든 수준결합에서 관측도수들의 분포는 곱다항분포를 따르게 된다.

$$\prod_{j=1}^{3}\prod_{i=1}^{5}\left\{\frac{n_{ij}!}{n_{ij1}!\,n_{ij2}!\,n_{ij3}!}\ \pi_{1|ij}^{\,n_{ij1}}\ \pi_{2|ij}^{\,n_{ij2}}\ \pi_{3|ij}^{\,n_{ij3}}\right\} \tag{3.3}$$

식 (3.3)에 누적로짓 혼합모형식 (3.2)를 적용하면 우도함수는

$$\prod_{j=1}^{3}\prod_{i=1}^{5}\left\{\frac{n_{ij}!}{n_{ij1}!\,n_{ij2}!\,n_{ij3}!}\ \left\{\frac{\exp(\alpha_1-\lambda_1 u_i-\sigma_B z_j)}{1+\exp(\alpha_1-\lambda_1 u_i-\sigma_B z_j)}\right\}^{n_{ij1}}\right. \tag{3.4}$$

$$\left\{\frac{\exp(\alpha_2-\lambda_2 u_i-\sigma_B z_j)}{1+\exp(\alpha_2-\lambda_2 u_i-\sigma_B z_j)}-\frac{\exp(\alpha_1-\lambda_1 u_i-\sigma_B z_j)}{1+\exp(\alpha_1-\lambda_1 u_i-\sigma_B z_j)}\right\}^{n_{ij2}}$$
$$\left\{1-\frac{\exp(\alpha_2-\lambda_2 u_i-\sigma_B z_j)}{1+\exp(\alpha_2-\lambda_2 u_i-\sigma_B z_j)}\right\}^{n_{ij3}}\Bigg\}$$

이다. 모수들의 최우추정값은 우도함수를 확률효과들인 시험장 수준효과들에 대하여 적분하여 주변우도함수를 구한다. 이 주변우도함수를 대수변환한 후 미지모수들에 대해 편미분하여 연립방정식들을 구한다. 이들 연립방정식들의 해는 Nelder and Mead(1965)의 심플렉스 방법을 이용하여 얻어진다. 구해진 해는 다음과 같다.

$\widehat{\alpha_1}$=0.4774(0.00012), $\widehat{\alpha_2}$=-0.5865(0.00016), $\widehat{\lambda_1}$=0.1494(0.0000021),
$\widehat{\lambda_2}$=0.0297(0.000002) 이고 $\widehat{\sigma_B}$=0.0000015(0.000002) 이다.

괄호안은 추정량의 분산에 대한 추정값을 나타내고 있다. 누적로짓 혼합모형식 (3.1)의 적합성을 알아보기 위한 측도로써 이용되는 이탈도의 값은 132.8이고 해당하는 자유도는 25이다. 평균이탈도가 1로부터 상당히 떨어져 있으므로 여러 다양한 모형을 적합시켜 자료에 적합한 모형을 살펴볼 수 있겠으나 순서형 다가자료에 대한 누적로짓 혼합모형을 적합시키는 방법을 제공하는 데 의미를 두고 있다.

4. 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 다른 한 요인은 반응변수에 영향을 미칠 수 있는 확률요인이다. 확률요인의 고려는 동질적인 실험단위들의 집단에서 임의로 일부 수준을 추출함으로써 발생하는 확률효과를 고려하고 있다. 다가의 순서형 반응범주들을 갖는 범주형 자료에 대한 변환으로 누적확률을 이용한 누적로짓 혼합모형을 제시하고 모형내 미지모수들의 추정값과 추정오차를 구하는 방법을 논의하였다.

참고문헌

1. 최재성(2004). A Mixed model for ordered response categories, 한국데이터정보과학회지, 제15권, 2호, 339-345.
2. Abramowitz, M. and Stegun, I.(1972). Handbook of mathematical functions, p. 924, Dover Publications, New York.
3. Agresti, Alan.(1990). Categorical data analysis, John Wiley and Sons, Inc., New York.

4. Im, S. and Gianola, D.(1988). Mixed models for binomial data with an application to lamb mortality, Applied Statistics, Vol. 37, 196-204.
5. McCullagh, P. and Nelder, J. A.(1989) Generalized linear models (2nd edition). Chapman and Hall, London.
6. Hosmer, W. David, and Lemeshow, Stanley(2000). Applied logistic regression (2nd edition), John Wiley and Sons, Inc., New York.

[2006년 1월 접수, 2006년 2월 채택]