Proceedings of Joint Conference of Korean Data And Information Science Society and The Korean Data Analysis Society, April 30 - May 1, 2004, pp. 121-126

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 there are some factors are fixed and others are random. Random factors are assumed to be coming from a two-way nested design for choosing individuals or experimental units to apply treatments. Estimation procedure for the unknown parameters in a suggested model is also discussed by an illustrated example.

Keywords : cumulative logit, mixed model, nested design, ordered data.

1. 서론

관측연구 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 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 두가지가 있다. 하나는 고정요인이고 다른 하나는 확률요인이다. 고정요인들은 대개 반응에 영향을 미치는 처치들이고 이들은 고정효과를 갖는 것으로 간주된다. 실험 또는 관측연구에서 확률요인들로 간주되는 변수들의 수준들은 수준들의 모집단에서 확률표본으로 추출된 일부 수준들임을 의미하고 있다. 따라서 표본으로 다른 수준들이 추출되면 개체의 반응에 영향을 미치는 효과는 달라 지게 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

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

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 관측연구로부터 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 자료분석방법은 모형에 근거하지 않은 자료분석방법 보다 체계적이고 효과적인 자료분석방법을 제공하게 된다. 따라서 자료분석을 위한 모형은 반응에 영향을 미치는 요인들의 유형에 따라 세가지로 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 표본추출계획과 관련된 확률요인과 처치로서의 고정요인을 고려한 혼합효과모형에서의 자료분석방법을 논의하고자 한다. 또한 순서형자료를 고려하고 있기 때문에 수집된자료는 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 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 표본추출단계로 부터 개체들의 반응에 영향을 미치는 확률요인들이 발생하게 될 때 혼합모형의 제시와 함께 모형내 미지모수들을 추론하는 방법을 논의한다.

2. 모형의 가정

개체의 반응과 관련한 몇가지 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 변

환을 가정한다. 셋째로 확률요인들은 실험의 설계구조나 개체들의 표본추출계획으로부터 발생하는 요인들 임을 가정한다. 확률요인은 개체의 반응에 영향을 미치는 독립변수로 유한개의 수준만이 실험에 고려되고 이 유한개의 수준들이 모집단으로부터 임의로 추출되었다고 가정하기 때문에 그 효과들은 확률효과들로 간주된다. 이는 확률요인이 반응을 나타내는 실험단위 또는 개체들에 행해졌음을 의미한다. 이러한 확률요인은 실험단위들의 표본추출방법으로부터 발생할 수 있다. 이러한 가정과 관련한 혼합모형의 제시를 위하여 다음과 같은 실험환경을 가정한다. 혼합모형의 의미는 처치구조에서 고정효과를 갖는 일부 고정요인들이 있고 하나 이상의 분산성분을 갖는 모형을 의미한다. 개체 또는 실험단위의 반응이 순서형의 다범주(multi-category)로 관측되고 각 반응범주의 확률에 영향을 미치는 독립변수로 세개의 요인 A, B와 C를 고려한다. 요인 A는 $i=1,2,\ldots,a$ 개의 수준들로 이루어진 고정요인(fixed factor)이고 요인 B와 C는 표본추출방법과 관련하여 발생하는 확률요인들로 가정한다. 요인 B는 $j=1,2,\ldots,b$개의 수준들로 주어지고 요인 C는 $k=1,2,\ldots,c$개의 수준들로 이루어진다. 개체에 대한 반응은 $t=1,2,\ldots,l$ 개의 순서형 범주들로 주어지는 반응변수 Y로 나타낸다.

연구자의 관심모집단에서 개체의 네가지 특성 A, B, C와 Y에 대한 조사는 네변수 A, B, C와 Y의 결합확률분포로 표현되거나 요인 A, 요인 B 그리고 요인 C의 주어진 수준하에서 조건부 확률분포로도 표현될 수 있다. 즉, $\{\pi_{t|ijk}\}$이다. 순서형 반응변수 Y의 각 범주에 속할 확률에 영향을 미치는 세 요인들의 효과를 추론하기 위하여 실험을 행한후 자료를 수집한다. 관심모집단에서 자료수집을 위해 먼저 지역들의 집단에서 일정크기의 지역 b개를 선정한후 선정된 지역 j내 구역집단에서 임의로 c개의 구역을 선정하는 이원지분계획을 이용하여 개체들을 추출한다 하자. 여기서 개별지역은 몇 개의 구역들로 이루어진다고 가정한다. 선정된 구역 k내의 모든 개체들 또는 일부개체들이 표본으로 추출되어 자료가 수집된다. 이러한 표본추출계획으로 인하여 개체의 반응확률에 영향을 미치는 두 가지 요인들이 발생하게 된다. 하나는 반응확률에 있어서 지역간의 변동을 나타내는 지역요인 과 다른 하나는 지역내 구역간의 변동을 나타내는 구역요인이다. 이들 요인들의 수준이 임의로 추출되기 때문에 확률요인들로 간주된다. 따라서, 개체의 반응에 영향을 미치는 요인들의 효과는 고정요인의 수준이 반응에 영향을 미치는 고정효과와 두 확률요인들의 수준들이 반응에 영향을 미치는 확률효과를 생각할 수 있다. 순서형 반응변수 Y의 관심범주들에 속할 확률들은 이들 요인들의 효과의 정도를 파악하기 위한 혼합모형으로 주어진다. 모형제시를 위한 다원분류표에서 n_{ijk}를 지역 j내 구역 k에서 요인A의 처치 또는 수준 i 가 행해진 실험단위들의 수라 두자. 따라서 n_{ijk}개 실험단위에서 순서형 반응변수 Y의 l 개 범주들의 관측도수는 n_{ijkt}로 주어지고 $\{n_{ijkt}\}$는 다항분포를 따르게 된다. 자료분석 모형들의 비교를 위해 반응변수가 명목형이고 고정요인이 또한 명목형 변수일 때를 생각해 보기로 한다. 이때의 자료모형은 다음과 같이 기술된다.

$$g(P(Y=t|ijk))=\alpha_t+\beta_{it}^{A}+\beta_{jt}^{B}+\beta_{kt}^{C}+\beta_{ijt}^{AB}+\beta_{ikt}^{AC} \tag{2.1}$$

$$i=1,2,\ldots,a,\ j=1,2,\ldots b,\ k=1,2,\ldots,c,\ t=1,2,\ldots,l-1.$$

여기서 $g(\cdot)$는 연결함수이고, α_t는 반응변수 Y가 범주 t로 반응할 때의 절편을 나타내며 $\{\beta_{it}^{A}\}$는 요인 A의 고정효과를 $\{\beta_{jt}^{B}\}$, $\{\beta_{kt}^{C}\}$, $\{\beta_{ijt}^{AB}\}$와 $\{\beta_{ikt}^{AC}\}$는 각기 요인 B와 C 그리고 교호작용 AB와 AC의 확률효과를 나타낸다. 요인 B와 C는 지분관계이므로 이들 간의 교호작용은 없다고 가정한다.

이들 확률효과들은 각기 $N(0, \sigma_B^2)$, $N(0, \sigma_C^2)$, $N(0, \sigma_{AB}^2)$와 $N(0, \sigma_{AC}^2)$을 따른다고 가정한다. 다음으로 순서형 자료를 분석하기 위한 모형은 고정요인을 측정척도에 따른 범주형 변수로 분류할 때 몇가지 가능한 모형들을 기술할 수 있다. 첫번 째는 요인 A가 측정측도로 분류할 때 명목형 변수일 때의 모형이다.

$$g(P(Y=t|ijk))=\alpha_t+\mu_i+\beta_{jt}^{B}+\beta_{kt}^{C(B)}+(\mu\beta)_{ijt}^{AB}+(\mu\beta)_{ikt}^{AC(B)} \tag{2.2}$$

$i=1,2,\ldots,a,\ j=1,2,\ldots b,\ k=1,2,\ldots,l-1$. 단 μ_i 는 요인 A의 수준효과를 나타낸다. 두번 째는 요인 A가 순서형 변수일 때의 모형이다.

$$g(P(Y=t|ijk))=\alpha_t+\lambda u_i+\beta_{jt}^{B}+\beta_{kt}^{C(B)} \tag{2.3}$$
$$i=1,2,\ldots,a,\ j=1,2,\ldots b,\ k=1,2,\ldots,l-1.$$

단, $\{u_i\}$는 요인 A의 수준들과 동일한 순서를 갖는 단조점수들이고 λ 는 연결함수 g로 변환된 값들에 대하여 u_i 에 따른 기울기이다. 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 누적로짓을 이용한 혼합효과 모형을 자료에 적합시켜 보고자 한다. 누적로짓은 다음과 같이 정의한다.

$$L_t=\log it[F_t(\boldsymbol{x})]=\log\frac{F_t(\boldsymbol{x})}{1-F_t(\boldsymbol{x})},\ k=1,2,\ \ldots\ l-1.$$

단, $F_t(\boldsymbol{x})=\pi_1(\boldsymbol{x})+\ \ldots\ +\pi_t(\boldsymbol{x})$인 범주 t까지의 누적확률을 나타낸다. $t=1,2,\ \ldots\ l$.

누적로짓을 이용할 때 식(2.3)은

$$L_{t|ijk}=\alpha_t+\lambda u_i+\beta_{jt}^{B}+\beta_{kt}^{C(B)},\ t=1,2,\ \ldots\ ,l-1. \tag{2.4}$$

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

$$L_{tijk} - L_{ti'j'k'} = \alpha_t + \lambda u_i + \beta_{jt}^{B} + \beta_{kt}^{C(B)} - \alpha_t - \lambda u_{i'} - \beta_{j't}^{B} - \beta_{k't}^{C(B)}$$
$$= \lambda(u_i - u_{i'}) + (\beta_{jt}^{B} - \beta_{j't}^{B}) + (\beta_{kt}^{C(B)} - \beta_{k't}^{C(B)})$$

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

두 누적로짓의 차이가 로그누적승산비임을 감안할 때, 로그누적승산비는 단순히 고정요인 A의 두 수준간에 효과차를 나타내고 있다. 또한 식(2.4)는 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

$$L_{tijk} = \alpha_t - \lambda u_i - \beta_{jt}^{B} - \beta_{kt}^{C(B)},\ t = 1, 2, \ \dots\ , l-1 \tag{2.5}$$

으로 주어진다.

3. 자료 예

촛불집회의 TV 보도에 대한 국민들의 반응을 조사하기 위해 지역집단에서 일부 지역을 임의로 추출하고 추출된 지역내 일부 기초자치단체 구역을 선정한다. 선정된 기초자치단체내 일부 시민을 추출한 다음 반응을 조사한다고 가정하자. 이때 뉴스보도에 대한 시각은 상당히 부정적인 시각, 객관적이고 사실적인 시각, 편파적이라고도 볼 수 있으나 긍정적인 시각의 세 범주로 관측된다. 시민들의 반응과 관련한 자료수집을 위해 전국 광역자치단체에서 임의로 3개 광역자치단체를 추출한다. 추출된 광역자치단체내에서 2개 기초자치단체를 임의로 추출한다. 선정된 기초자치단체에서 일부시민을 임의로 추출하여 반응을 조사하게 된다. 이때 시민들의 세 개 반응범주에 속할 확률은 광역자치단체간의 변동에 따른 영향을 받을 수 있고 또한 광역자치단체내 지역자치단체간의 변동에 따른 영향을 예상할 수 있다. 반응범주의 확률에 영향을 미칠 수 있는 고정요인으로 촛불집회와 관련한 언론매체의 접촉회수를 고려한다. <표 3.1>은 이들 세 개 요인을 고려한 생성자료표이다. 생성자료표의 자료를 분석하기 위한 모형으로 식(2.5)를 이용하기로 한다. 광역자치단체 j, 기초자치단체 k 그리고 접촉회수 i에서 각 범주내 관측도수를 n_{ijkt} 라 두면 세 요인의 모든 수준 결합에서 관측도수들의 분포는 곱다항분포를 따르게 된다.

$$\prod_{j=1}^{3}\prod_{k=1}^{2}\prod_{t=1}^{4}\left\{\frac{n_{ijk}!}{n_{ijk1}!\,n_{ijk2}!\,n_{ijk3}!}\ \pi_{1|ijk}^{n_{ijk1}}\ \pi_{2|ijk}^{n_{ijk2}}\ \pi_{3|ijk}^{n_{ijk3}}\right\} \tag{3.1}$$

식(3.1)에 누적로짓 혼합모형식(2.5)를 적용하면 우도함수는

$$\prod_{j=1}^{3}\prod_{k=1}^{2}\prod_{i=1}^{4}\left\{\frac{n_{ijk}!}{n_{ijk1}!\,n_{ijk2}!\,n_{ijk3}!}\left\{\frac{\exp(\alpha_1-\lambda_1 u_i-\beta_{j1}^{B}-\beta_{k1}^{C(B)})}{1+\exp(\alpha_1-\lambda_1 u_i-\beta_{j1}^{B}-\beta_{k1}^{C(B)})}\right\}^{n_{ijk1}}\right. \tag{3.2}$$

$$\left\{1-\frac{\exp(\alpha_1-\lambda_1 u_i-\beta_{j1}^{B}-\beta_{k1}^{C(B)})}{1+\exp(\alpha_1-\lambda_1 u_i-\beta_{j1}^{B}-\beta_{k1}^{C(B)})}-\frac{1}{1+\exp(\alpha_2-\lambda_2 u_i-\beta_{j2}^{B}-\beta_{k2}^{C(B)})}\right\}^{n_{ijk2}}$$

$$\left.\left\{\frac{1}{1+\exp(\alpha_2-\lambda_2 u_i-\beta_{j2}^{B}-\beta_{k2}^{C(B)})}\right\}^{n_{ijk3}}\right\}$$

이다. 모수들의 최우추정값은 우도함수를 확률효과들인 광역자치단체의 수준들과 관련된 효과들과 기초자치단체들의 수준과 관련한 효과들에 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.

<표 3.1> 촛불집회의 생성자료

광역자치단체(B)	기초자치단체(C)	접촉회수(A)	반응(Y) 부정적	중립적	긍정적
서울	중구	0	20	30	50
		1	35	26	48
		2	22	47	65
		3	19	27	45
	강남구	0	24	31	60
		1	27	18	48
		2	15	47	48
		3	19	32	51
광주	광산구	0	45	22	18
		1	27	26	49
		2	52	17	24
		3	24	36	45
	동구	0	27	25	37
		1	25	36	52
		2	32	27	35
		3	32	18	55
대구	남구	0	43	22	36
		1	42	18	57
		2	38	16	38
		3	42	17	28
	수성구	0	36	22	48
		1	42	18	26
		2	54	20	19
		3	38	24	35

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. Hosmer, W. David, and Lemeshow, Stanley. (2000). *Applied logistic regression* (2nd edition), John Wiley and Sons, Inc., New York.