Journal of Korean
Data & Information Science Society
2006, Vol. 17, No. 1, pp. 187 ~ 193

A Continuation-Ratio Logits Mixed Model for Structured Polytomous Data

Jaesung Choi[1)]

Abstract

This paper shows how to use continuation-ratio logits for the analysis of structured polytomous data. Here, response categories are considered to have a nested binary structure. Thus, conditionally nested binary random variables can be defined in each step. Two types of factors are considered as independent variables affecting response probabilities. For the purpose of analyzing categorical data with binary nested strutures a continuation-ratio mixed model is suggested. Estimation procedure for the unknown parameters in a suggested model is also discussed in detail by an example.

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

1. 서론

분석하고자 하는 자료가 다원분류표상의 범주형 자료일 때, 반응변수의 관측값은 일정순서를 갖는 다가의 범주로 주어지고 다수의 독립변수들이 고려된 경우의 자료분석 모형을 논의해 보기로 한다. 반응변수의 다가범주들이 자연스러운 순서를 가질 때, 순서형 반응범주들에 적용할 수 있는 다양한 변환을 생각할 수 있으나 본 논문에서는 연속비 로짓(continuation-ratio logit)을 이용한 로짓변환에 대한 자료분석모형을 논의하고자 한다. 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 연속비 로짓을 이용한 경우의 모형에 관한 논의는 찾아보기가 쉽지 않다. 개체 또는 실험단위의 반응에 대한 단순척도(pure scale)의 관측반응이 다가의 범주로 주어질 때, 실험 또는 조사로부터 수집된 자료는 다가

1) Professor, Department of Statistics, Keimyung University, 1000 Sindang-Dong, Daegu 704-701, Korea
E-Mail : jschoi@kmu.ac.kr

자료를 구성하게 되고 이들이 반응범주의 도수로 표현되면 다항자료(multinomial data)라 한다. 그러나, 순서형 반응범주에 연속비 로짓변환을 고려하는경우에 단순척도의 관측범주들이 아닌 반응범주가 어떤 구조적 특성을 갖는 경우의 적용을 생각해 보기로 한다. 따라서, 반응범주들의 구조적 특성으로 복합척도(complex scale)에 의한 순서형 반응들로 이루어지는 범주들을 고려한다. 복합척도의 순서형 반응범주를 위한 다양한 척도들을 고려할 수 있으나 여기서는 단순한 이가의 지분구조를 생각해 보기로 한다. 반응범주의 확률에 영향을 미치는 독립변수들로 실험 또는 관측조사의 단계에서 서로 다른 독립변수들로 구성되는 경우를 고려한다.

2. 모형의 가정

연속비 로짓을 이용한 모형설정을 위해 다음과 같은 실험환경을 생각해 보기로 하자. 우선 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 연속비 로짓 변환을 가정한다. 셋째로 반응의 관측범주는 이가의 지분구조로 구성된다. 넷째로 독립변수들은 개별 단계에서 서로 다른 변수들로 이루어진다. 다단계의 실험 또는 관측조사로부터 주어지는 개체의 반응범주들은 대개 이가의 지분구조를 갖는 범주들로 구성된다. 마지막으로, 독립변수중 한 변수는 확률변수임을 가정한다. 예로써 여러 지역에 제품 생산공장을 갖고 있는 어느 회사가 자사의 생산제품에 대한 시장의 신뢰도를 높이기 위해 세 단계의 품질관리의 공정단계를 거쳐서 상품을 출하시킨다고 가정하자. 생산제품의 불량률을 줄이기 위한 일차검사는 전자시스템에 의한 성능테스트가 행해진다. 일차검사의 성능테스트에 합격한 제품에 한하여 이차공정단계에는 충격시험에 의한 내구성검사로 제품을 조사한다. 이차검사를 통과한 제품은 외장검사를 통하여 결함여부를 결정한다고 하자. 생산제품의 품질관리와 관련된 결과의 반응범주는 다음과 같이 요약될 수 있다.

A_1은 일차검사에서 불합격,
B_1는 일차검사에서 합격하였으나 이차에서 불합격,
C_1은 이차검사에서 합격하였으나 삼차에서 불합격, 그리고
C_2는 삼차에서 합격한 범주이다.

이들 네 범주들은 검사단계별 순서가 정해진 순서범주들로 간주된다. 또한, 각 단계에서 서로 다른 특성 즉, 독립변수들로 합격여부가 결정되고 있음을 알 수 있다. 그리고, 매 단계에서의 개체에 대한 관측여부는 합격이거나 불합격의 이가 반응을 나타내고 있으므로 반응범주들은 이가의 지분구조를 갖게 된다. 각 범주에서의 확률을 다음과 같이 정의한다.

$$\pi_1 = P(A_1),\ \pi_2 = P(B_1),\ \pi_3 = P(C_1)\ \text{이고}\ \ \pi_4 = P(C_2)\text{이다.}$$

단, $\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$이다.

상호배반인 네 범주의 이러한 확률정의는 각 단계에서의 합격률과 관련된 연속비 로짓을 정의하기 위함이다. 검사과정에서 고려된 독립변수들을 x_1, x_2, 그리고 x_3라 두자. 각 반응범주확률은 이들 변수들의 함수로 표현될 수 있다. 주어진 $\boldsymbol{x}=(x_1, x_2, x_3)$의 한 값에서 반응범주가 넷이므로 세 개의 연속비 로짓을 정의한다. 즉,

$$L_1 = \log\left(\frac{\pi_1(\boldsymbol{x})}{\pi_2(\boldsymbol{x})+\pi_3(\boldsymbol{x})+\pi_4(\boldsymbol{x})}\right), \text{ 이고} \tag{2.1}$$

$$L_2 = \log\left(\frac{\pi_2(\boldsymbol{x})}{\pi_3(\boldsymbol{x})+\pi_4(\boldsymbol{x})}\right)$$

$$L_3 = \log\left(\frac{\pi_3(\boldsymbol{x})}{\pi_4(\boldsymbol{x})}\right) \text{ 이다.}$$

반응범주들이 이가의 지분구조로 이루어져 있기 때문에 각 범주의 확률을 각 단계에서의 조건부 확률로 표시해 보기로 한다. 우선, 범주 j의 조건부 확률을

$$\rho_j(\boldsymbol{x}) = \frac{\pi_j(\boldsymbol{x})}{\pi_j(\boldsymbol{x})+\cdots+\pi_4(\boldsymbol{x})}, \quad j=1,2,3.$$

으로 정의한다. $\rho_j(\boldsymbol{x})$는 반응이 범주 j를 포함한 그 이상의 범주에 속한다는 조건이 주어졌을 때, 범주 j에 속할 확률을 나타내고 있다. 앞서 정의된 세 연속비 로짓을 조건부 확률들의 로짓으로 변환해 보면 다음과 같다.

$$L_j = \log\left(\frac{\rho_j(\boldsymbol{x})}{1-\rho_j(\boldsymbol{x})}\right), \quad j=1,2,3.$$

다시말하면, 세 연속비 로짓은 각 검사단계에서의 조건부 불합격률의 승산에 대한 대수변환으로 나타낼 수 있음을 보여주고 있다. 따라서, 이가지분구조의 순서형 다가범주에 대한 연속비 로짓에 대한 모형은

$$L_j = \log\left(\frac{\rho_j(\boldsymbol{x})}{1-\rho_j(\boldsymbol{x})}\right) = \alpha_j + \boldsymbol{\beta}'\boldsymbol{x}, \quad j=1,2,3. \tag{2.2}$$

로 주어진다. 즉,

$$\begin{aligned} L_1 &= \alpha_1 + \boldsymbol{\beta}'\boldsymbol{x}, \\ L_2 &= \alpha_2 + \boldsymbol{\beta}'\boldsymbol{x}, \text{ 이고} \\ L_3 &= \alpha_3 + \boldsymbol{\beta}'\boldsymbol{x} \end{aligned} \tag{2.3}$$

이다. 이 모형은 고려된 독립변수들의 모든 수준에서 관측되는 세 개의 연속비 로짓

에 대해 독립변수들의 효과를 파악하기 위한 일반적인 모형이다. 그러나, 가정에서 실험 또는 관측조사의 각 단계에서 제품을 검사하는 변수들이 다르므로 모형 (2.3)의 이용은 적합하지 않게 된다. 검사의 매 단계에서 제품의 합격여부와 관련된 변수는 단일의 x_i, $i=1,2,3$, 이다. 따라서 적용하고자 하는 모형은 x_i의 한 관측값에서

$$L_j = \alpha_j + \beta_j x_i, \quad j=1,2,3 \tag{2.4}$$

로 주어지게 된다.

3. 모수의 추론

식 (2.4)를 이용한 모형내 모수들의 추론방법을 생각해 보기로 한다. 표본으로 생산공장들의 모집단에서 세 개의 공장을 표본으로 추출한 다음 선정된 공장, k, $k=1,\cdots,3$ 에서 크기 n_k인 제품을 추출하였을 때, 고정벡타 $\boldsymbol{x}$ 의 값에서 각 범주에 속하는 개체의 수를 n_{jk}, $j=1,\cdots,4$라 두자. $\{n_{jk}, j=1,\cdots,4\}$는 다음의 다항분포를 따른다.

$$\frac{n_k!}{n_{1k}!\cdots n_{4k}!} \quad \pi_{1k}^{\,n_{1k}} \cdots \pi_{4k}^{\,n_{4k}}$$

위의 다항분포는 각 단계에서 정의된 조건부 확률로 나타낼 때, 다음과 같이 표현된다.

$$\left\{\frac{n_k!}{n_{1k}!(n_k-n_{1k}!)}(\rho_{1k}(x_1))^{n_{1k}}(1-\rho_{1k}(x_1))^{n_k-n_{1k}}\right\}\cdots \tag{3.1}$$
$$\left\{\frac{(n_k-n_{1k}-n_{2k})!}{n_{3k}!(n_k-n_{1k}-n_{2k}-n_{3k})!}(\rho_{3k}(x_3))^{n_{3k}}(1-\rho_{3k}(x_3))^{n_k-n_{1k}-n_{2k}-n_{3k}}\right\}$$

즉, 다항분포는 각 검사단계에서 제품이 조사된다는 조건하에서, 검사받는 제품이 불합격할 확률이 조건부 확률인 $\rho_j(\boldsymbol{x})$, $j=1,2,3$, 을 미지모수로 갖는 이항분포들의 곱으로 표현될 수 있음을 보여주고 있다. 따라서 전체의 우도함수는 벡타 $\boldsymbol{x}$의 모든 값에서 이들 확률들의 곱인 곱다항분포로 주어진다. 식 (3.1)에 식 (2.4)를 대입하면

$$\left\{\frac{n_k!}{n_{1k}!(n_k-n_{1k}!)}\left(\frac{\exp(\alpha_1+\beta_1 x_1+l_k)}{1+\exp(\alpha_1+\beta_1 x_1+l_k)}\right)^{n_{1k}}\left(\frac{1}{\exp(\alpha_1+\beta_1 x_1+l_k)}\right)^{n_k-n_{1k}}\right\}\cdots \tag{3.2}$$
$$\left\{\frac{(n_k-n_{1k}-n_{2k})!}{n_{3k}!(n_k-n_{1k}-n_{2k}-n_{3k})!}\left(\frac{\exp(\alpha_3+\beta_3 x_3+l_k)}{1+\exp(\alpha_3+\beta_3 x_3+l_k)}\right)^{n_{3k}}\right\}$$
$$\left\{\left(\frac{1}{1+\exp(\alpha_3+\beta_3 x_3+l_k)}\right)^{n_k-n_{1k}-n_{2k}-n_{3k}}\right\}$$

를 구한다. 모형내 미지모수들을 추론하기 위하여 전체 우도함수를 대수변환하여 미지모수들에 편미분하여 얻어진 연립방정식들을 풀면 최우추정치가 구해진다. 그러나 구해진 편미분 방정식들이 미지모수들에 대해 비선형이므로 수치연산의 알고리즘을 이용할 수 있다. 확률효과가 포함되지 않은 고정 모형인 경우에는 다양한 연산프로그램을 이용할 수 있다. 모형내에 확률효과가 포함된 혼합효과 모형인 경우에는 변환로짓의 유형에 따라 제한적으로 이용가능한 상업용 프로그램이 있으나 대부분의 경우 모수추정치를 구하기 위한 프로그램을 마련하는 것이 필요하다. 그러나, 반응구조가 이가의 지분구조를 갖는 경우에 변환로짓으로 연속비 로짓을 이용할 때, sas macro를 이용할 수 있다.

4. 생산제품의 검사자료 예

연속비 로짓을 이용한 모형내 미지모수들을 추론하기 위한 방법을 설명하기 위한 예를 생각해 보기로 한다. 국내 여러지역에 생산공장을 보유하고 있는 업체 A는 자사에서 생산되고 있는 제품의 신뢰도를 확보하기 위한 방편으로 불량률을 줄이기 위해 시장에 출하하기 전에 세 단계의 제품검사를 실시한다고 하자. 세 단계의 제품검사는 성능테스트, 충격테스트 그리고 외장결함테스트로 이루어진다. 생산공장집단에서 임의로 세 개 공장을 표본으로 추출한 다음 추출된 공장에서 각기 100개의 제품을 표본으로 추출하여 다음 자료를 얻었다고 하자. <표 4.1>의 검사자료는 각 단계에서 생산제품의 불량률에 영향을 미치는 독립변수는 각각 수준이 셋이고 서로 다른 척도의 변수들 임을 나타내고 있다. 예로써, 일차검사의 성능테스트에서 수준 50에서 검사받은 100개의 제품중 8개는 불량품임을 나타내며, 이보다 높은 수준의 60에서 검사받은 제품들은 수준 50에서 불량품 8개를 제외한 92개중 7개가 불량품임을 나타내고 있다. 그리고 수준 70에서의 5개의 불량품은 15개를 제외한 85개의 조사받은 제품중 불량개수이다. 일차검사에서 합격하지 않은 제품들은 다음 이차검사단계에서 검사받지 않게 된다.

<표 4.1> 제품의 검사자료

생산공장	일 차 검 사		이 차 검 사		삼 차 검 사	
	x_1	불량개수	x_2	불량개수	x_3	불량개수
1	50	8	1	4	1	1
	60	7	2	5	2	2
	70	5	3	1	3	1
2	50	10	1	3	1	2
	60	5	2	8	2	0
	70	5	3	4	3	1
3	50	4	1	6	1	1
	60	8	2	5	2	0
	70	6	3	6	3	2

위 자료를 분석하기 위하여 식 (2.4)의 모형을 적합시켜 보기로 한다.

$$\begin{aligned} L_1 &= \alpha_1 + \beta_1 x_1, \\ L_2 &= \alpha_2 + \beta_2 x_2, \text{ 이고} \\ L_3 &= \alpha_3 + \beta_3 x_3 \end{aligned} \tag{4.1}$$

식 (4.1)은 표본을 취하기 전 모집단에 대한 모형이므로 표본추출후의 자료분석모형은 다음과 같이 주어진다.

$$\begin{aligned} L_{1k} &= \alpha_1 + \beta_1 x_1 + l_k, \\ L_{2k} &= \alpha_2 + \beta_2 x_2 + l_k, \text{ 이고} \\ L_{3k} &= \alpha_3 + \beta_3 x_3 + l_k \end{aligned} \tag{4.2}$$

식 (4.2)에서 l_k, $k=1,\cdots,3$, 는 확률효과를 나타내므로 다음의 표준정규변수들의 값으로 치환된 모형식을 이용함이 편리하다.

$$\begin{aligned} L_{1k} &= \alpha_1 + \beta_1 x_1 + \sigma_z z_k, \\ L_{2k} &= \alpha_2 + \beta_2 x_2 + \sigma_z z_k, \text{ 이고} \\ L_{3k} &= \alpha_3 + \beta_3 x_3 + \sigma_z z_k \end{aligned} \tag{4.3}$$

이다. 검사자료표의 자료를 분석하기 위한 모형으로 식 (4.3)을 이용하기로 한다. 표본으로 추출된 세 개의 생산공장에서 조사된 제품들의 관측도수들에 대한 결합확률분포는 다음의 곱다항분포를 따르게 된다.

$$\prod_{k=1}^{3} \left\{ \frac{n_k!}{n_{1k}!\cdots!n_{4k}!} \quad \pi_{1k}^{\ n_{1k}} \cdots \ \pi_{4k}^{\ n_{4k}} \right\} \tag{4.4}$$

식 (4.4)에 식 (4.3)을 대입하면 우도함수는

$$\begin{aligned} &\prod_{k=1}^{3} \left\{ \frac{n_k!}{n_{1k}!n_{1k}'!} \left\{ \left(\frac{\exp(\alpha_1 + \beta_1 x_1 + \sigma_z z_k)}{1+\exp(\alpha_1 + \beta_1 x_1 + \sigma_z z_k)} \right)^{n_{1k}} \left(\frac{1}{1+\exp(\alpha_1 + \beta_1 x_1 + \sigma_z z_k)} \right)^{n_{1k}'} \right\} \right. \\ &\left. \left\{ \frac{n_{1k}'!}{n_{2k}!n_{2k}'!} \left(\frac{\exp(\alpha_2 + \beta_2 x_2 + \sigma_z z_k)}{1+\exp(\alpha_2 + \beta_2 x_2 + \sigma_z z_k)} \right)^{n_{2k}} \left(\frac{1}{1+\exp(\alpha_2 + \beta_2 x_2 + \sigma_z z_k)} \right)^{n_{2k}'} \right\} \right\} \\ &\left\{ \frac{n_{2k}'!}{n_{3k}!n_{4k}!} \left\{ \left(\frac{\exp(\alpha_3 + \beta_3 x + \sigma_z z_k)}{1+\exp(\alpha_3 + \beta_3 x + \sigma_z z_k)} \right)^{n_{3k}} \left(\frac{1}{1+\exp(\alpha_3 + \beta_3 x + \sigma_z z_k)} \right)^{n_{4k}} \right\} \right. \end{aligned} \tag{4.5}$$

이다. 단, $n_{1k}' = n_k - n_{1k}$, $n_{2k}' = n_k - n_{1k} - n_{2k}$ 이고 $n_{4k} = n_k - n_{1k} - n_{2k} - n_{3k}$이다. 위의 우도함수를 대수변환하여 로그우도함수를 구한다. 공장효과를 나타내는 표준정규변수 z_k에 대하여 적분하면 주변우도함수를 구할 수 있다. 구해진 주변우도함수를 미지모수들에 대해 편미분하여 얻은 방정식들을 연립으로 풀면 최우추정치를 구할 수 있다. 이들 연립방정식들의 해는 다음과 같이 얻어진다.

$\widehat{\alpha_1}$=-2.0848(0.8885), $\widehat{\alpha_2}$=-2.5333(0.3898), $\widehat{\alpha_3}$=-4.1374(0.7505)
$\widehat{\beta_1}$=-0.00863(0.01492), $\widehat{\beta_2}$=-0.2217(0.1950), $\widehat{\beta_3}$=0.01994(0.3475)

이다. 생산공장에 따른 변동효과는 거의 없는 것으로 추정된다.

괄호안은 추정량의 표준편차에 대한 추정값을 나타내고 있다. 연속비 로짓 혼합모형식 (4.3)의 적합성을 알아보기 위한 측도로써 이용되는 이탈도의 값은 18.5825이고 해당하는 자유도는 19이다. 평균이탈도가 1과 상당히 근사하므로 자료분석에 대한 모형으로 타당하다고 판단할 수 있겠으나 이보다는 순서형 다가자료에 대한 연속비 로짓 혼합모형을 적합시키는 방법의 제공에 의미를 두고 있다.

4. 결론

본 논문은 범주형 자료가 다단계의 실험 또는 관측조사로부터 수집되는 순서형의 다가자료를 분석하기 위한 혼합모형을 논의하고 있다. 다단계의 실험 또는 조사를 통하여 관측되는 반응범주들은 일정한 순서를 갖게 되고 반응범주들은 각 단계에서의 지분변수와 관련된 반응구조로 표현될 수 있음에 착안하고 있다. 모형설정을 위하여 이가의 지분구조로 이루어진 다가의 순서형 반응범주들을 가정하였다. 이러한 가정하에 가능한 변환으로 연속비 로짓변환을 이용할 수 있음을 나타내고 있다. 또한, 실험의 각 단계에서 반응에 영향을 줄 수 있는 서로 다른 독립변수들을 고려할 때의 모집단 모형을 언급하고 있다. 실제 자료분석에 이용되는 표본모형은 관측단위들의 표본추출과 관련된 확률요인을 추가함으로서 혼합모형으로 표현됨을 알 수 있다. 따라서 반응범주가 이가의 지분구조를 갖는 다가의 순서형 범주들을 연속비 로짓 변환한 자료에 대해 연속비 로짓 혼합모형을 제시하고 모형내 미지모수들의 추정값과 추정오차를 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.
5. 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]