응용통계연구 제 **14**권 *1*호, *2001*년, *pp. 103–110*

복합척도의 범주형 자료에 대한 연속모형*

최재성[1)]

요 약

본 논문은 다단계로 구성되는 실험 또는 조사를 통한 개체의 반응척도가 계층적 구조를 갖는 복합척도의 범주형 자료를 나타낼 때, 이들 자료를 분석하기 위한 모형으로 조건부 지분변수를 이용한 연속모형을 제시하고 있다.

주요용어: 계층적구조, 복합척도, 조건부 지분변수.

1. 서론

실험 또는 조사로 부터 수집되는 자료가 복합척도의 범주형 자료일 때, 자료분석 모형은 자료의 성격, 구조 및 특성에 따라 다양한 모형들로 제시될 수 있다. 개체의 복합반응척도에 대한 범주형 자료들은 반응변수들 간에 구조적 특성을 갖지 않는 경우와 구조적 특성을 고려해야 하는 경우로 분류할 수 있다. 본 연구의 내용은 복합반응척도에 의해 주어지는 자료들이 조사 또는 실험의 특성상 구조적으로 조건부 지분변수를 활용할 수 있는 범주형 자료일 때, 이들 자료를 분석하기 위한 모형제시에 관심을 두고 있다. 복합반응척도란 개체의 다변량 반응들에 대한 척도들이 서로 다른 척도들로 나타남을 의미한다. 예를들면, 이변량 반응에서 한 반응은 순서척도에 의해 관측되고 다른 반응은 명목척도로 관측됨을 말한다. 또 다른 유형의 복합반응척도는 다변량의 반응들이 지분구조를 가질 때, 복합반응척도를 갖는다고 말한다. 개체에 대한 반응들이 복합척도이고, 이들 반응에 대한 자료들이 범주형 자료일 때, 수집된 자료를 분석하기 위한 모형에 관한 논의는 여러 문헌에서 제시되고 있다. Agresti(1990)(269쪽)는 복합척도의 관측자료를 분석하기 위한 대수선형모형들에 관하여 구체적으로 논의하고 있다. Anderson and Aitkin(1985)은 조사설문지의 분석을 위한 면담자들의 변동을 다루었다. 이들은 이원지분 조사계획으로 발생하는 분산성분들을 추정하기 위하여 로지트 모형에 대한 최우추정법을 제시했다. Im and Gianola(1988)는 이원지분배열로 부터 수집된 이항자료에 대한 혼합모형에서 최우추정치를 계산하기 위한 심플렉스 방법을 논의했다. Cox and Snell(1989)은 반응에 있어서 계층적 구조를 갖는 다가자료의 분석을 위하여 조건부 지분 확률변수의 이용에 관하여 논의했다. 또한, McCullagh and Nelder(1989)는 지분구조를 갖는 다가자료에 대해 가능한 모형들을 제시하고 있다. 확률효과를 갖는 이가자료를 분석하기 위하여 Conaway(1990)는 각 개체에 대한 반복측정치들이 독립임을 의미하는 국부적 독립모형과 반응들 간의 추가종속성을 반영하기 위한 종속모수

* 본 연구는 1999년도 계명대학교 비사연구기금으로 이루어졌음.

1) (704-701) 대구광역시 달서구 신당동 1000, 계명대학교 통계학과, 교수
E-mail: jschoi@kmucc.keimyung.ac.kr

들을 포함하는 종속모형들을 다루었다. 본 연구는 앞서 언급된 문헌들과는 달리 조사 또는 실험이 다단계의 과정으로 행해지고, 각 단계에서 개체에 대한 반응척도가 서로 다른 단계에서의 반응척도와 동일한 척도를 갖거나 또는 다른 척도로 관측될 수 있으나, 실험 또는 조사가 다단계로 구성되어 있는 특성으로 인하여 각 단계에서의 반응들은 계층적 구조를 갖는 복합척도의 자료분석에 관심을 갖는다.

2. 모형

복합척도를 나타내는 범주형 반응변수들간의 관련성을 규명하기 위한 구조적 모형으로 다양한 유형의 대수선형모형을 생각해 볼 수 있겠다. 그러나 반응간에 계층적 구조를 갖고, 각 반응들에 영향을 미치는 독립변수들이 같지 않을 때, 자료분석에 이용될 모형은 반응간의 계층적 구조와 각 단계의 반응에 영향을 미치는 설명변수들을 고려한 모형이 적절하다. 이러한 모형을 설명하기 위한 예로 자동차 운전면허시험을 생각해 보자. 운전면허시험은 세 단계의 평가시험들로 구성된다. 첫 단계는 도로교통법과 관련된 필기시험이며, 두 번째 단계는 필기시험에 합격한 사람에 한하여 응시할 수 있는 장내기능시험이고, 세 번째 단계는 장내기능시험에 합격한 후 응시할 수 있는 도로주행시험이다. 따라서, 운전면허시험에 응시한 응시자들의 집단에서 면허시험에 대한 결과들은 다음과 같이 네개의 상호배반인 범주로 주어진다. 즉,

필기시험에 불합격인 범주, 장내기능시험에 불합격인 범주, 도로주행시험에 불합격인 범주, 그리고 도로주행시험에 합격인 네 범주이다.

이들 네 범주들은 계층적 구조를 갖고 있음을 알 수 있다. 즉, 장내기능시험에 불합격으로 주어지는 범주는 첫 단계에서의 반응범주가 필기시험에서의 합격임을 전제로 하고 있다. 도로주행시험에서의 합격과 불합격은 첫 단계의 필기시험과 두 번째 단계의 장내기능시험에 모두 합격한 경우에만 관측되는 범주이다. 각 단계에서 관측되는 반응에 영향을 미칠 수 있는 독립변수들을 고려할 때, 이들 독립변수들이 동일한 변수들로만 이루어지지 않음을 알 수 있다. 예를 들면, 필기시험결과에 영향을 미치는 변수들로 필기시험을 준비한 기간, 응시자의 연령등을 고려할 수 있다. 장내기능시험의 결과에 영향을 미치는 변수들로는 연습시간, 면허종류, 그리고 성별등을 생각할 수 있다. 도로주행시험의 결과와 관련된 변수들로는 도로주행 연수시간, 응시회수, 그리고 운전경력등을 고려할 수 있겠다. 운전면허시험의 결과로 주어지는 네범주의 계층적 구조와 각 단계에서 고려되는 독립변수들을 포함시키는 모형을 고려하기 위하여 다음과 같이 변수들을 정의한다. Y_1은 단계1에서 필기시험에 합격이면 1 아니면 0의 값을 취하는 이가변수로 정의한다. $Y_{2|1}$은 단계2에서 장내기능시험에 합격이면 1 아니면 0의 값을 취하는 조건부 이가변수로 정의된다. 왜냐하면, 장내기능시험에 응시할 수 있는 자격은 필기시험에 합격해야 하기 때문이다. 즉, $Y_1 = 1$이 주어졌을 때, $Y_{2|1}$은 0 또는 1의 값을 취할 수 있기 때문이다. 또한, $Y_{3|11}$은 단계3에서 도로주행시험에 합격이면 1 아니면 0의 값을 취하는 조건부 이가변수로 정의된다. 각 단계에서 정의된 이가의 반응변수들은 서로 다른 유형의 시험에 따른 반응변수들이므로 이들 반응에 영향을 미치는 각 단계에서의 독립변수들도 일반적으로 같지 않음을 알 수 있다. 따라서, 네

범주의 계층적 구조를 고려하고 각 단계에서 정의된 반응변수들에 영향을 미치는 서로 다른 독립변수들을 고려한 모형은 다음과 같이 주어진다.

$$\begin{aligned} g(P(Y_1 = 1)) &= \alpha_1 + \sum_{i=1}^{p} \beta_i x_i \\ g(P(Y_{2|1} = 1)) &= \alpha_2 + \sum_{j=1}^{r} \theta_j z_j \\ g(P(Y_{3|11} = 1)) &= \alpha_3 + \sum_{k=1}^{s} \delta_k t_k \end{aligned} \tag{2.1}$$

단, g는 연결함수(link function)이고, p개의 x, r개의 z, 그리고 s개의 t는 각 단계의 반응에 영향을 미치는 독립변수들을 나타낸다. 다가의 범주형 자료(polytomous categorical data)가 계층적 구조를 갖는 복합척도의 자료이고, 각 단계에서 고려되는 독립변수들이 같지 않을 때, 자료분석을 위한 모형으로 연속모형을 이용하는 것이 적절하다. 왜냐하면, 각 단계에서 정의된 반응변수들은 서로 다른 관측결과와 관련된 변수이고, 이들 반응변수에 영향을 미치는 독립변수들도 서로 다르므로 개별적인 모형하에 관련변수들의 효과들을 추론하는 것이 체계적이고 효율적이기 때문이다.

3. 모수의 추론

모형내 모수들의 추론을 위하여 면허시험 응시자의 집단에서 크기 n인 표본을 추출할 때, 복합척도의 네범주에 속하는 관측도수의 분포는 다항분포를 따르게 된다. 이때, 다항분포가 조건부 지분변수들의 이항분포들의 곱으로 표현될 수 있음을 나타내기 위하여 독립변수들의 각 수준결합에서 각 범주에 속할 확률과 도수를 다음과 같이 정의한다.

필기시험에 합격할 범주에 속할 확률과 도수를 각각 π_1, n_1이라 두자. 장내기능시험에 합격할 범주에 속할 확률과 도수를 각각 π_{11}, n_{11}이라 두자. 도로주행시험에 합격할 범주에 속할 확률과 도수를 각각 π_{111}, n_{111}이라 두자.

따라서, 다항표본추출로 부터 크기 n인 표본을 취했을 때, 각 범주에 속하는 도수들의 확률분포는 다음과 같은 다항분포를 따르게 된다.

$$\begin{aligned} &\frac{n!}{(n-n_1)!(n_1-n_{11})!(n_{11}-n_{111})!n_{111}!}(1-\pi_1)^{n-n_1}(\pi_1-\pi_{11})^{n_1-n_{11}} \\ &(\pi_{11}-\pi_{111})^{n_{11}-n_{111}}\pi_{111}^{n_{111}} \end{aligned} \tag{3.1}$$

위의 다항분포는 각 단계에서 정의된 조건부 지분변수를 이용할 때, 다음과 같이 이항분포들의 곱으로 표시된다.

$$\begin{aligned} &\frac{n!}{n_1!(n-n_1)!}\pi_1{}^{n_1}(1-\pi_1)^{n-n_1}\frac{n_1!}{n_{11}!(n_1-n_{11})!}\left(\frac{\pi_{11}}{\pi_1}\right)^{n_{11}}\left(1-\frac{\pi_{11}}{\pi_1}\right)^{n_1-n_{11}} \\ &\frac{n_{11}!}{n_{111}!(n_{11}-n_{111})!}\left(\frac{\pi_{111}}{\pi_{11}}\right)^{n_{111}}\left(1-\frac{\pi_{111}}{\pi_{11}}\right)^{n_{11}-n_{111}} \end{aligned} \tag{3.2}$$

다가의 범주형 반응변수에 대한 척도가 계층적 구조의 이가 반응을 나타내는 복합척도일 때, 이러한 구조적 특성을 갖는 다가자료를 분석하기 위한 표본분포는 다항분포이나 각

단계에서 정의된 이가의 조건부 지분변수의 관심범주에 대한 도수들의 분포가 이항분포를 따르므로 이들 분포의 곱으로 표시됨을 알 수 있다. 연구자의 관심이 각 단계에서 관측되는 합격률과 각 단계에서 이들 관심확률에 영향을 미칠 수 있는 독립변수들과의 관련성을 파악하고자 하는 경우에 단순히 운전자의 면허시험에서 관측되는 상호배반인 네범주의 확률과 도수분포로 주어진 다항분포는 의미가 없다. 그러나 이 분포는 각 단계에서 정의된 조건부 지분변수의 정의로 부터 관심확률과 관심도수의 다항분포로 표현됨으로써 모수추정에 이용된다. 연속모형내 포함된 모수들의 최우추정값들은 독립변수들의 모든 수준결합에서 관측되는 다항분포들의 곱으로 표현되는 우도함수를 미지모수들에 관하여 미분하여 얻어진 연립방정식들을 해결하여 구한다. 그러나 추정방정식들이 미지모수들간에 비선형이므로 Nelder and Mead(1965)의 심플렉스 방법을 이용한다.

4. 운전면허시험의 예

어느 대학에 입학한 신입생중 운전면허시험에 응시한 학생들을 대상으로 조사한 자료의 결과표가 다음과 같다고 하자.

표 4.1: 운전면허시험의 생성자료

학습기간	장내운전	도로주행	시험결과			
(일)	연습시간	연수시간	필기불합격	장내불합격	주행불합격	주행합격
7	15	12	6	2	1	1
		15	5	3	0	2
	20	12	8	3	2	2
		15	9	3	1	2
10	15	12	14	18	4	4
		15	18	14	2	6
	20	12	18	7	5	5
		15	17	7	4	7
15	15	12	4	6	2	2
		15	3	5	1	2
	20	12	4	4	3	2
		15	4	3	1	4

위의 자료를 분석하기 위한 모형으로 다음 연속모형들을 가정한다.

$$\begin{aligned} logit(P(Y_1 = 1)) &= logit(p_1) = \alpha_1 + \beta_1 x \\ logit(P(Y_{2|1} = 1)) &= logit(p_2) = \alpha_2 + \beta_2 z \\ logit(P(Y_{3|11} = 1)) &= logit(p_3) = \alpha_3 + \beta_3 t \end{aligned} \tag{4.1}$$

단, logit은 로지트 연결함수(link function)이고, x는 필기시험을 위한 학습기간, z는 장내 기능시험을 위한 운전연습시간이고 t는 도로주행시험을 위한 연수시간을 나타낸다. 또한, $p_1 = \pi_1$, $p_2 = \frac{\pi_{11}}{\pi_1}$ 이고 $p_3 = \frac{\pi_{111}}{\pi_{11}}$으로 정의한다. 모형(4.1)에 근거한 다항분포(3.2)의 형태는 다음과 같다.

$$\begin{aligned}
&\frac{n!}{n_1!(n-n_1)!}\left(\frac{\exp(\alpha_1+\beta_1 x)}{1+\exp(\alpha_1+\beta_1 x)}\right)^{n_1}\left(1-\frac{\exp(\alpha_1+\beta_1 x)}{1+\exp(\alpha_1+\beta_1 x)}\right)^{n-n_1}\\
&\frac{n_1!}{n_{11}!(n_1-n_{11})!}\left(\frac{\exp(\alpha_2+\beta_2 z)}{1+\exp(\alpha_2+\beta_2 z)}\right)^{n_{11}}\left(1-\frac{\exp(\alpha_2+\beta_2 z)}{1+\exp(\alpha_2+\beta_2 z)}\right)^{n_1-n_{11}} \qquad (4.2)\\
&\frac{n_{11}!}{n_{111}!(n_{11}-n_{111})!}\left(\frac{\exp(\alpha_3+\beta_3 t)}{1+\exp(\alpha_3+\beta_3 t)}\right)^{n_{111}}\left(1-\frac{\exp(\alpha_3+\beta_3 t)}{1+\exp(\alpha_3+\beta_3 t)}\right)^{n_{11}-n_{111}}
\end{aligned}$$

연속모형의 가정하에 모형내 미지모수들을 추정하기 위한 과정을 살펴보기 위하여 표본자료에 대한 우도함수를 LH라 두자. 이때, 우도함수는 다음과 같이 구해진다.

$$\begin{aligned}
LH \;=\; &\prod_{i=1}^{3}\prod_{j=1}^{2}\prod_{k=1}^{2}\left\{\frac{n_{ijk}!}{n_{1ijk}!(n_{ijk}-n_{1ijk})!}\right.\\
&{p_{1ijk}}^{n_{1ijk}}(1-p_{1ijk})^{n_{ijk}-n_{1ijk}}\}\\
&\left\{\frac{n_{1ijk}!}{n_{11ijk}!(n_{1ijk}-n_{11ijk})!}(p_{2ijk})^{n_{11ijk}}(1-p_{2ijk})^{n_{1ijk}-n_{11ijk}}\right\} \qquad (4.3)\\
&\left\{\frac{n_{11ijk}!}{n_{111ijk}!(n_{11ijk}-n_{111ijk})!}(p_{3ijk})^{n_{111ijk}}(1-p_{3ijk})^{n_{11ijk}-n_{111ijk}}\right\}
\end{aligned}$$

단, i는 x의 3수준, 즉, $x_1 = 7$, $x_2 = 10$, $x_3 = 15$를 나타내고, j는 z의 2수준, k는 t의 2수준을 나타낸다. 식(4.2)에서 처럼 식(4.3)의 p들을 식(4.1)의 해당하는 모수들로 표현된 확률을 대입한후 대수변환을 취하면 대수우도함수를 얻을 수 있다. 대수우도함수를 LLH로 나타낼 때, LLH를 미지모수들에 편미분하여 얻은 방정식들은 미지모수들에 관하여 비선형이기 때문에 최우추정치들을 얻기 위한 numerical algorithm으로 심플렉스 방법을 이용한다. 그리고 모형의 적합성을 판단하기 위한 이탈도는 현재모형(current model)이 포화모형(saturated model)으로 부터의 변이정도를 나타내는 측도로써 우도비의 대수변환값을 -2배한 값으로 정의된다. 정의에 따른 이탈도를 계산하기 위하여 최우추정치에 의한 LLH를 $L\hat{L}H$라 두고 포화모형에 의한 LLH를 $L\hat{L}H_f$라 둘 때, 이탈도의 값은 $-2[L\hat{L}H - L\hat{L}H_f]$으로 주어지고 이들 값은 근사적으로 카이제곱분포를 따르게 된다. 이탈도 정의에서 기술된 포화모형이란 관측치의 개수만큼 많은 미지모수들을 포함하고 있는 모형을 의미한다. 위 자료에 해당하는 포화모형은 관측치의 수가 36개이므로 이들 수 만큼 미지모수를 포함하고 있는 모형을 말한다.

세 단계의 시험으로 구성되는 운전면허시험의 결과로서 관측되는 복합척도의 네 반응범주에 영향을 미칠 수 있는 독립변수들과의 관련성을 알아보기 위한 모형은 각 단계에서 행해지는 시험결과와 관련된 관심범주의 확률에 영향을 미칠 수 있는 연속적인 모형하에서 자료분석하는 것이 타당함을 보여주고 있다. 왜냐하면, 연구자의 관심이 면허취득과 관련된 각 단계에서 행해진 합격률에 있기 때문에 필기시험에서 고려된 학습기간은 단계2에

서 행해진 장내기능시험의 결과와 전혀 관련없는 변수이고, 장내기능시험과 관련된 변수는 도로주행시험과 무관한 변수로 간주되기 때문이다. 최우추정법에 의해 추정된 모형내 미지모수들의 최우추정치들은 다음과 같다.

$$\begin{aligned} \hat{\alpha_1} &= -1.041(0.0973), \quad \hat{\beta_1} = 0.125(0.0090), \\ \hat{\alpha_2} &= -1.246(0.1490), \quad \hat{\beta_2} = 0.082(0.0108), \\ \hat{\alpha_3} &= 0.478(0.2711), \quad \hat{\beta_3} = -0.004(0.0141). \end{aligned}$$

괄호안은 추정치의 표준오차를 나타내고 있다. 자료분석을 위한 연속모형(4.1)의 적합성을 알아보기 위한 측도로써 이용되는 이탈도의 값은 47.57이고 해당하는 자유도는 30이므로 자료를 잘 설명하고 있다고 볼 수 있다. 위의 자료분석과 관련된 모형설정과정을 설명하기 위하여 장내기능시험과 관련된 장내운전 연습시간이 도로주행시험의 합격률에도 영향을 미칠 수 있는 독립변수임을 가정해보자. 이 경우의 모형은

$$\begin{aligned} &logit(P(Y_1 = 1)) = logit(p_1) = \alpha_1 + \beta_1 x \\ &logit(P(Y_{2|1} = 1)) = logit(p_2) = \alpha_2 + \beta_2 z \\ &logit(P(Y_{3|11} = 1)) = logit(p_3) = \alpha_3 + \beta_3 z + \beta_4 t \end{aligned} \tag{4.4}$$

로 주어짐을 알 수 있다. 모형(4.4)의 가정하에 주어진 이탈도의 값은 모형(4.1)에서의 이탈도와 동일한 값으로 관측된다. 이는 실제적으로 장내운전연습시간이 도로주행시험의 합격률에 거의 영향을 미치지 않는 변수임을 의미하고 있다. 모형(4.4)과 관련된 모수 추정치들은 다음과 같다.

$$\begin{aligned} \hat{\alpha_1} &= -1.036(0.0741), \quad \hat{\beta_1} = 0.124(0.0644), \\ \hat{\alpha_2} &= -1.146(0.1285), \quad \hat{\beta_2} = 0.075(0.0111), \\ \hat{\alpha_3} &= 0.403(0.1027), \quad \hat{\beta_3} = 0.022(0.0670), \\ & \qquad\qquad\quad \hat{\beta_4} = -0.017(0.0585). \end{aligned}$$

다양한 특성의 자료를 분석하기 위한 모형의 한 예로써, 최재성(1996)은 질병발생집단에서 관심질병의 처치효과들을 효과적으로 분석하기 위하여 감염율을 고려한 모형을 제시하고 있다.

5. 결론

실험 또는 조사를 통하여 얻게되는 다가의 범주형 자료를 분석하기 위한 모형이 다수의 반응변수들을 포함하고 각 반응변수에 영향을 미치는 독립변수들이 별개의 모형에서 고려되어야 하는 경우 단일의 대수선형모형은 자료분석에 적합하지 않게 된다. 따라서, 본 논문은 개체의 반응척도가 계층적 구조의 복합척도이고, 조사의 각 단계에서 반응에 영향을 미치는 변수들이 동일하지 않을 때 각 단계에서 조건부 지분변수를 정의함으로서 자료분석에 필요한 모형을 얻을 수 있음을 구체적인 예를 통하여 제시 하고 있다. 모형설정과정의 예에서 알 수 있듯이 연구자의 관심이 각 단계에서 정의된 이가 지분변수의 성공확률에 있

을 때, 제시된 모형내 모수를 추론하기 위한 표본분포로써 다항분포가 각 단계에서의 성공확률과 도수로써 표현되기 위하여 상호배반인 네 범주의 확률을 주의깊게 정의하여야 함을 나타내고 있다.

참고문헌

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

[2] Anderson, D.A. and Aitkin, M. (1985). Variance component models with binary response: Interviwer variability, *Journal of the Royal Statistical Society Series B*, Vol. **47**, 203-210.

[3] Conaway, M.R. (1990). A random effects model for binary data. *Biometrics*, Vol. **46**, 317-328.

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

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

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

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

[8] 최재성. (1996). 질병의 범주적 자료에 대한 통계적 분석모형, <응용통계연구>, 제9권 1호, 1-15.

[2000년 3월 접수, 2001년 1월 채택]

A Sequence of Models for Categorical Data with Compound Scales*

Jaesung Choi[1)]

ABSTRACT

This paper considers a multistage experiment. Response scales can be same or different from stage to stage. When variables are of nested structure, the response variable at each stage can be defined conditionally. For analysing such data with compound scales, this paper suggests a sequnce of dependence models and shows how to set up a sequence of models for the driver's liscense test data.

Keywords: Nested structure; Compound scales; Conditional nested variables.

* The present research has been conducted by the Bisa Research Grant of Keimyung University in 1999.

1) Professor, Department of Statistics, Keimyung University.
E-mail: jschoi@kmucc.keimyung.ac.kr