한국통계학회논문집 제4권 2호
1997년 8월 pp. 377 - 384

지분구조의 다가자료에 관한 모형

최 재 성[1)]

요 약

본 논문은 지분구조를 갖는 범주형 자료가 명목상의 다가자료일 때, 지분구조의 각 단계에서 정의될 수 있는 지분변수들의 유형 과 지분변수들의 관심확률들에 영향을 미치는 변수들을 고려한 자료분석 모형들을 제시하고 있다.

1. 서론

조사, 실험 또는 관측연구로 부터 수집되는 자료들이 범주형 자료들이라 하자. 이들 범주형 자료들은 반응변수들의 관측되는 범주의 수에 따라 이가(binary) 반응변수(response variable)의 관측자료일 때는 이가자료(binary data) 그리고 다가(polytomous) 반응변수의 관측자료일 때는 다가자료(polytomous data)로 구분되어 진다. 또한 이들 자료가 관측범주의 도수로 주어지면 이가자료는 이항자료(binomial data)로 다가자료는 다항자료(multinomial data)로 불리어 진다. 이가 또는 이항자료는 단일 반응변수에 대한 관측자료이기 때문에 반응척도(response scale)에 있어서 지분구조를 갖지 않으나 다가자료 또는 다항자료의 경우 측정척도(measurement scale)가 복합척도(compound scale)의 자료일 수 있다.

복합척도의 다가자료에 대한 자료분석을 위하여 Cox 와 Snell(1989)은 다가변수들을 조건부로 정의된 이가변수들의 지분구조(nested structure)로 표현하는 것이 바람직하다고 제의했다. McCullagh 와 Nelder(1989)는 관측반응들의 지분구조 또는 계층구조(hierarchical structure)를 이용한 연속모형들에 관해 논의하고 있다. 최재성(1996)은 조사질병의 한 예방백신의 효과가 질병발생집단의 감염율을 고려하였을 때 어떻게 영향을 받는 가를 알아보기 위한 모형제시에서 다가자료의 지분구조를 이용하고 있다. 이들 문헌들은 범주형 자료가 다가자료이고 지분구조를 갖는 명목상 자료(nominal data)일 때, 조건부 지분 이가변수를 활용한 모형을 논의하고 있다. 그러나 실험, 조사 또는 관측연구로 부터 수집되는 자료들이 반드시 조건부 지분 이가변수들이 정의될 수 있는 지분구조의 명목상 자료로만 주어지지 않기 때문에 이와는 다른 유형의 지분구조를 갖는 다가자료들을 분석하기 위한 모형의 제시는 자료의 효과적인 분석을 위하여 필요함을 알 수 있다.

1) (704-701) 대구 광역시 달시구 신당동 계명대학교 자연과학대학 통계학과 부교수

2. 지분구조

복합척도의 범주형 자료가 지분구조를 갖는 다가자료일 때, 지분구조의 특성에 따른 조건부 지분변수들을 정의할 수 있다. 지분구조의 각 단계에서 정의된 조건부 지분변수가 취하는 값들의 확률에 대한 일반화 선형 모형(generalized linear model)을 생각할 수 있다. 지분구조의 각 단계에서 고려된 일반화 선형모형들의 집단은 지분구조를 갖는 관심의 다가자료를 분석하기 위하여 이용할 수 있는 일련의 종속모형들 임을 알 수 있다. 이러한 종속모형에 관한 논의는 Conaway(1990)에서도 제시되나 이는 개체별 반복측정한 자료를 분석하기 위한 종속모형이다. 다가자료의 지분구조를 활용하여 자료분석에 이용될 수 있는 일련의 종속모형들은 지분구조의 각 단계에서 정의되는 조건부 변수들에 의해 결정된다.

2.1 이가의 지분변수

지분구조의 반응척도가 지분구조의 각 단계에서 이가의 지분변수들 만이 정의될 수 있는 경우를 생각해 보기로 한다. 이가의 지분변수들이 정의될 수 있는 예로써 세 단계의 수준으로 구성되어 있는 어떤 어린이용 전자오락을 생각해 보자. 이 때 두 번째 단계의 수준에서 전자오락을 즐기기 위해서는 첫 번째 단계의 수준을 통과하였을 때만 가능하고, 세 번째 단계의 수준에서 전자오락을 즐기기 위해서는 앞서의 두 단계의 수준들에서 통과하였을 때만 가능하다고 하자. 각 단계에서의 관측반응은 통과하거나 실패하거나 이다. 따라서 이 실험의 모든 가능한 결과들의 집합인 표본공간, S, 는 각 단계에서 통과하였으면 P 실패하였으면 N으로 나타낼 때 다음과 같이 주어진다.

$$S=\{N,(P,N),(P,P,N),(P,P,P)\}$$

즉, 이 실험에서의 표본공간은 상호배반인 네 개의 범주로 주어지고 이들 범주들은 다음과 같다.

$A_1=\{N\}$, $A_2=\{(P,N)\}$, $A_3=\{(P,P,N)\}$ 그리고 $A_4=\{(P,P,P)\}$ 이다.

$A_2 \cup A_3 \cup A_4$는 첫 번째 단계의 수준을 통과한 범주를 나타내고, $A_3 \cup A_4$는 두 번째 단계의 수준을 통과한 범주를 나타내므로 반응범주들은 지분구조를 가짐을 알 수 있다. 왜냐하면, A_2는 첫 단계에서의 결과가 통과일 때만 관측되고, A_3 와 A_4는 두 번째 단계에서의 결과가 통과일 때만 관측될 수 있기 때문이다.

관측범주들의 지분구조를 이용할 때 표본공간에서 정의될 수 있는 변수들은 다음과 같다.

Y_1은 첫 단계에서의 결과가 P일 때 1의 값을 취하고 N일 때 0의 값을 취하는 이가변수라 정의한다. Y_2는 첫 단계에서의 결과가 P일 때, 두 번째 단계에서의 결과가 P이면 1 N이면 0의 값을 취하는 이가변수라 정의한다. 그리고 Y_3는 두 번째 단계에서의 결과가 P일 때 세 번째 단계에서의 결과가 P이면 1 N이면 0의 값을 취하는 변수로 정의한다. 따라서 Y_2의 값들은 Y_1이 1일 때만 관측되고, Y_3의 값들은 Y_1이 1이고, Y_2가 1일 때만 관측되므로 Y_2와 Y_3는 조건

부 지분 이가변수들 이다. 표본공간에서 정의된 세 변수들의 표본점들은 다음과 같다.

$$W=\{(y_1,y_2,y_3):(0,\phi,\phi),(1,0,\phi),(1,1,0),(1,1,1)\}$$

단, $(0,\phi,\phi)$는 Y_1이 0이기 때문에 Y_2와 Y_3가 정의되지 않는 표본점을 나타내고, $(1,0,\phi)$는 Y_2가 0이기 때문에 Y_3가 정의되지 않는 표본점을 나타낸다. 따라서 표본점들의 표기의 편의상 $(0,\phi,\phi)$를 $(0,0,0)$로 $(1,0,\phi)$를 $(1,0,0)$로 표현하는 것이 바람직하다. 이 경우 표본점들의 집합은 다음과 같다.

$$V=\{(y_1,y_2,y_3):(0,0,0),(1,0,0),(1,1,0),(1,1,1)\}$$

2.1.1 모형

수집된 범주형 자료가 다가자료이고 2.1절에서의 예와같은 지분구조를 가질 때, 자료를 분석하기 위한 모형을 생각해 보기로 한다. 모형설정을 위하여 전자오락의 네 관측범주들에 대한 확률은 관측범주들 간의 지분구조를 이용하여 다음과 같이 정의된 확률들을 이용하여 구할 수 있다.

$\pi_1=P(A_2\bigcup A_3\bigcup A_4)$,

$\pi_2=P(A_3\bigcup A_4)$, 그리고

$\pi_3=P(A_4)$ 이다.

여기서 π_1은 첫 단계의 수준을 통과할 확률이고 π_2는 두 번째 단계의 수준을 통과할 확률이며 π_3는 세 번째 단계의 수준을 통과할 확률이다. 다시말하면 전자오락의 세 단계에서의 수준들은 서로 다른 난이도로 구성되어 있음을 의미한다. 위의 확률로 부터

$P(A_1)=1-\pi_1$, $P(A_2)=\pi_1-\pi_2$, 이고 $P(A_3)=\pi_2-\pi_3$ 임을 알 수 있다.

이 전자오락을 고안한 프로그래머는 시판에 앞서 어린이들에게 충분히 흥미를 유발시킬 수 있도록 각 단계의 수준이 적당한 난이도를 갖는 가에 관심을 가질 수 있다. 따라서 각 범주에서의 관측확률뿐만 아니라 각 지분구조의 단계에서 주어진 수준을 통과할 확률도 주요 관심문제가 될 것이다. 각 단계의 수준을 통과할 확률에 영향을 미치는 변수들로 이 전자오락을 할 수 있는 어린이의 성별, 연령, 그리고 각 단계의 수준에서 주어지는 방해물의 갯수등이 고려될 수 있다. 이들 설명변수를 고려한 모형은 2.1절에서 정의된 조건부 지분 이가변수들로 표현할 때 다음과 같다.

$$g(\pi_1)=g[P(Y_1=1)]=\alpha_1+\boldsymbol{\beta}'\boldsymbol{x}$$

$$g(\pi_2/\pi_1)=g[P(Y_2=1|y_1=1)]=\alpha_2+\boldsymbol{\beta}'\boldsymbol{x}$$

$$g(\pi_3/\pi_2)=g[P(Y_3=1|y_1=1,y_2=1)]=\alpha_3+\boldsymbol{\beta}'\boldsymbol{x}$$

단, α_1, α_2, 와 α_3는 각 선형예측의 절편이고 $\boldsymbol{\beta}$는 회귀모수들의 벡타이며 $\boldsymbol{x}$는 설명변수들의 벡타이다.

2.2 다가의 지분변수

지분구조를 갖는 범주형 자료로서 지분구조의 각 단계에서 정의되는 변수들이 모두 다가인 경우를 생각해 보기로 한다. 다가의 지분변수들이 정의될 수 있는 경우를 설명하기 위하여 다음과 같은 관측조사를 행한다고 가정하자. B형 간염백신을 제조 판매하고자 하는 한 제약회사가 간염백신의 세 가지 제조방법중 어느 방법이 효과적인 가를 알아보기 위하여 초등학교 1학년에 재학중인 어린이들의 집단에서 적절한 표본추출방법으로 표본을 추출하여 추출된 어린이들을 대상으로 세 종류의 백신, B_1, B_2, 와 B_3, 에 의한 접종이 행해진다고 하자. 표본으로 추출된 어린이들에게 예방접종을 실시하기 위한 첫 번째 단계로써 혈액검사를 행할 수 있다. 혈액검사의 결과 어린이들은 이미 면역이 되어 있거나, 감염되었거나, 또는 미감염되었거나의 세 범주로 조사된다고 하자. 따라서 두 번째 단계는 감염되지 않은 어린이들을 대상으로 세 종류의 백신을 확률화(randomization)에 근거하여 예방접종 완료후 항체생성여부에 대한 조사를 할 수 있다. 백신종류별 항체생성 여부와 관련된 범주들은 B_1에 의해 항체가 생성되었거나, B_2에 의해 생성되었거나, B_3에 의해 생성되었거나, 그리고 항체가 생성되지 않았거나의 네개의 범주로 관측된다고 하자. 위와 같은 조사를 통하여 관측될 수 있는 상호배반인 범주들은 다음과 같다.

A_1은 조사시점에서 면역이 되어 있는 범주이고,

A_2는 조사시점에서 감염이 되어 있는 범주이며,

A_3는 조사시점에서 감염되어 있지않고 B_1에 의한 예방접종후 항체가 생성된 결과의 범주이고,

A_4는 조사시점에서 감염되어 있지않고 B_2에 의한 예방접종후 항체가 생성된 결과의 범주이며,

A_5는 조사시점에서 감염되어 있지않고 B_3에 의한 예방접종후 항체가 생성된 결과의 범주이고,

A_6는 조사시점에서 감염되어 있지않고 세 종류의 백신에 의한 예방접종후 항체가 생성되지 않은 결과의 범주를 나타낸다.

$A_3 \cup A_4 \cup A_5 \cup A_6$는 예방접종된 범주를 나타내고 있기 때문에 반응범주들 간에 지분구조를 갖고 있음을 알 수 있다. 수집된 자료가 지분구조를 갖는 다가자료이므로 지분구조의 특성을 이용한 모형설정을 위하여 조건부 지분 다가변수를 정의할 수 있다. 백신접종을 위한 혈액검사의 단계에서 관측되는 세개의 범주에 해당하는 수값을 취하는 변수를 Y_T라 두자. 이때 Y_T는 다음과 같이 정의된다.

$$Y_T=\begin{cases}0 & \text{혈액검사의 결과에 의해 면역이 되어 있다면}\\ 1 & \text{혈액검사의 결과에 의해 감염이 되어 있다면}\\ 2 & \text{혈액검사의 결과에 의해 감염이 되어 있지 않다면}\end{cases}$$

예방접종후 항체생성여부와 관련된 두 번째 단계의 반응범주들에 해당하는 수값을 취하는 조건부 지분변수를 Y_{NI}라 두자. 이때 Y_{NI}는 다음과 같이 정의할 수 있다.

$$Y_{NI}=\begin{cases}0 & \text{예방접종후 항체가 생성되지 않으면}\\ 1 & B_1\text{에 의해 항체가 생성되면}\\ 2 & B_2\text{에 의해 항체가 생성되면}\\ 3 & B_3\text{에 의해 항체가 생성되면}\end{cases}$$

위에서 정의된 Y_{NI}는 Y_T가 2의 값을 취할 때만 네개의 가능한 값중 한 값을 취할 수 있으므로 조건부 지분 다가변수임을 알 수 있다. 이들 두변수들의 정의로 부터 주어지는 표본점들의 집합은 다음과 같다.

$$W_1=\{(y_T,y_{NI}):(0,\phi),(1,\phi),(2,0),(2,1),(2,2),(2,3)\}$$

단, $(0,\phi)$는 Y_T가 0이기 때문에 Y_{NI}가 정의되지 않는 표본점이고, $(1,\phi)$는 Y_T가 1이기 때문에 Y_{NI}가 정의되지 않는 표본점이다. 다시말하면, $(0,\phi)$는 혈액검사의 결과 면역이 되어 있으므로 예방접종이 필요하지 않은 어린이들의 범주에 해당하는 표본점이고, $(1,\phi)$는 혈액검사의 결과 감염되어 있으므로 예방접종이 필요하지 않은 어린이들의 범주를 나타내는 표본점임을 의미하고 있다. 2.1절에서의 논의에서와 같이 표본점 $(0,\phi)$는 $(0,0)$로 $(1,\phi)$는 $(1,0)$로 표현하는 것이 바람직하다. 이 때 표본점들의 집합은 다음과 같이 주어진다.

$$V_1=\{(y_T,y_{NI}):(0,0),(1,0),(2,0),(2,1),(2,2),(2,3)\}$$

2.2.1 모형

자료분석을 위한 모형설정을 위하여 실험을 통하여 관측되는 범주들의 확률을 다음과 같이 정의한다.

$$\pi_I=P(A_2),\ \pi_{B1}=P(A_3),$$
$$\pi_{B2}=P(A_4),\ \pi_{B3}=P(A_5),\ \text{이고}\ \pi_{NI}=P(A_3\cup A_4\cup A_5\cup A_6)$$

라 두자. 위와 같은 확률의 정의로 부터 다른 근원사상들에 대한 확률은 다음과 같이 구해진다.

$P(A_1)=1-\pi_I-\pi_{NI}$ 이고 $P(A_6)=\pi_{NI}-\pi_{B1}-\pi_{B2}-\pi_{B3}$이다.

이 조사의 예에서 관심확률들은 관측되는 각 범주에 속할 개별확률이라기 보다는 오히려 지분구조의 각 단계에서 주어지는 범주들의 확률이 될 것이다. 따라서, 이들 범주들의 확률은 지분구조의 각 단계에서 정의된 확률변수들을 이용하여 표현될 수 있고, 지분 변수들의 반응에 영향을 미칠 수 있는 설명변수들의 효과를 알아보기 위한 모형은 다음과 같이 기술될 수 있다.

$$g[P(Y_T = y_T)] = \alpha_1 + \boldsymbol{\beta}'\boldsymbol{x},$$

$$g[P(Y_{NI} = y_{NI} | Y_T = 2)] = \alpha_2 + \boldsymbol{\beta}'\boldsymbol{x}$$

단, g는 $(0,1)$의 구간을 $(-\infty, +\infty)$의 구간으로 대응시키는 미분가능한 단조함수이고, α_1과 α_2는 절편이다. $\boldsymbol{\beta}$는 회귀모수들의 벡타이고, $\boldsymbol{x}$는 설명변수들의 벡타이다.

2.3 두 유형의 지분변수

각 연구분야들로 부터 수집되는 자료들이 지분구조를 갖는 범주형 자료라고 할지라도 이들 자료가 반드시 이가의 지분변수 또는 다가의 지분변수들로만 정의될 수 있는 자료로 주어진다고 보기는 어렵다. 따라서 이절에서는 두 유형의 지분변수들이 정의될 수 있는 예를 생각해 보기로 한다.

어느 교육학자가 어떤 조사지역에서 1997 학년도에 입학한 초등학교 일학년 학생들을 대상으로 입학전 유치원에서의 교육여부가 입학후 어떤 특정과목의 학업성취와 관련성이 있는 가를 알아보고자 한다고 가정하자. 이 때 조사자의 대상모집단에서 한 개체에 대한 조사는 다음 두 단계로 행해진다고 하자. 조사의 첫 단계는 입학전 유치원 교육을 받았는가 받지 않았는 가를 알아보는 것이다. 그 다음 두 번째 단계는 유치원 교육을 받은 학생일 때, 관심의 특정과목에서 학업성취도를 A, B, 그리고 C등급으로 분류하여 관측한다고 하자. 이러한 관측조사에서 모든 가능한 범주들은 다음과 같다.

E_1은 입학전 유치원 교육을 받지 않은 학생일 범주이고,

E_2는 입학전 유치원 교육을 받았고, 특정과목에서의 학업성취도가 A등급인 학생일 범주이며,

E_3는 입학전 유치원 교육을 받았고, 특정과목에서의 학업성취도가 B등급인 학생일 범주이며,

E_4는 입학전 유치원 교육을 받았고, 특정과목에서의 학업성취도가 C등급인 학생일 범주이다.

이 때, 교육학자의 관심은 첫 단계에서 조사된 유치원 교육의 유무와 관련하여 1997 학년도에 입학한 학생들의 유치원 교육을 받은 비율이 각 가정에서의 자녀수, 유치원 교육비용, 그리고 부모의 학력정도등의 변수들과 관련성이 있는 가를 알아보고자 할 수 있다. 또한 조사의 두 번째 단계에서 조사되는 관심의 특정과목에서의 성취도에 따른 비율들이 유치원에서의 주당 수업시간 수, 성별, 그리고 유치원에서 특정과목과 관련된 학과목에 대한 수업여부등의 변수들에 대한 효과를 파악하고자 할 수 있다. 관측조사에서 주어지는 반응범주들이 지분구조를 갖는 두 관심변수들의

관측범주들이므로 지분구조의 각 단계에서 정의될 수 있는 변수들은 다음과 같다.

$$Y_K=\begin{cases}0 & \text{유치원 교육을 받지 않은 학생이면}\\ 1 & \text{유치원 교육을 받은 학생이면}\end{cases},$$

$$Y_P=\begin{cases}1 & \text{특정과목의 성취도가 } A\text{등급이면}\\ 2 & \text{특정과목의 성취도가 } B\text{등급이면}\\ 3 & \text{특정과목의 성취도가 } C\text{등급이면}\end{cases}$$

위의 두 변수들의 정의로 부터 주어지는 표본점들의 집합, W_2,는 다음과 같다.

$$W_2=\{(y_K, y_P):(0,\phi),(1,1),(1,2),(1,3)\}$$

단, $(0,\phi)$는 Y_K가 0이기 때문에 Y_P가 정의되지 않는 표본점을 나타낸다. 다시말하면, 특정과목의 성취도는 유치원 교육을 받은 학생들에서만 조사되고 있음을 의미한다. 이 때, 표본공간의 표기의 편의상 $(0,\phi)$를 $(0,0)$로 표현하면 표본점들의 집합은 다음과 같다.

$$V_2=\{(y_K, y_P):(0,0),(1,1),(1,2),(1,3)\}$$

2.3.1 모형

지분구조의 각 단계에서 정의되는 변수들이 이가의 지분변수와 다가의 지분변수로 주어지는 관측조사의 예에서 수집되는 자료를 분석하기 위한 모형설정을 위하여 관심사상들의 확률을 다음과 같이 정의한다.

$\pi_A=P(E_2)$, $\pi_B=P(E_3)$, 이고 $\pi_K=P(E_2\cup E_3\cup E_4)$라 두자.

다른 근원사상들에 대한 확률들은 위의 확률의 정의로 부터 다음과 같이 구해진다.

$P(E_1)=1-\pi_K$이고, $P(E_4)=\pi_K-\pi_A-\pi_B$ 이다.

지분구조의 첫 단계에서 정의된 이가의 지분변수에 대한 관심비율은 유치원 교육을 받은 학생들의 비율, π_K, 이다. 지분구조의 두 번째 단계에서는 특정과목에 대한 성취도의 범주를 나타내는 조건부 지분 다가변수의 관측값들의 확률에 관심이 있으므로 지분구조의 각 단계에서 정의된 지분변수들의 확률로 모형을 기술하면 다음과 같다.

$$g[P(Y_K=y_K)]=g(\pi_K)=\alpha_1+\boldsymbol{\beta}_1'\boldsymbol{x}$$

$$g[P(Y_P=y_P|Y_K=1)]=\alpha_2+\boldsymbol{\beta}_2'\boldsymbol{x}$$

단, g는 연결함수이고, α_1과 α_2는 절편들이다. 그리고 $\boldsymbol{\beta}_1$과 $\boldsymbol{\beta}_2$는 회귀모수들의 벡타이고 각

선형예측에서의 설명변수들의 벡타 $\boldsymbol{x}$가 동일하지 않음을 나타내고 있다.

3. 결론

본 논문은 조사, 실험, 또는 관측연구를 통하여 수집되는 자료들이 지분구조의 범주형 자료들일 때, 이들 자료들을 분석하기 위한 모형들은 다양한 지분구조를 토대로 제시될 수 있음을 보여주고 있다. 다시 말하면, 계층구조 또는 지분구조를 갖는 반응척도가 지분구조의 각 단계에서 정의될 수 있는 지분변수들이 어떤 유형의 변수들인 가를 고려함으로써 자료분석에 유용한 모형들을 논의하고 있다. 지분구조를 갖는 범주형 자료에 대한 분석모형으로 지분구조를 고려한 모형은 지분구조를 고려하지 않은 모형에 의한 분석보다 추론에 있어서 효과적이고 체계적인 분석방법을 제공하게 된다. 왜냐하면, 자료의 수집방법, 실험의 단계, 또는 조사의 단계에서 주어지는 정보를 모형설정에 이용하여 관심변수들에 대해 세부적이고 구체적인 분석이 이루어 질 수 있기 때문이다.

참고문헌

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

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

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

[4] 최재성 (1996). 질병의 범주적 자료에 대한 통계적 분석모형, 「응용통계연구」, 제9권 1호, 1-15.