응용통계연구 제10권 2호
1997년 9월 pp. 325-338

범주형 자료에 대한 혼합모형[1)]

최 재 성[2)]

요 약

본 논문은 병후면역이 일생동안 지속되는 한 감염성 질병의 발생집단에서 질병발생 집단내 감염되지 않은 개체들에 대한 어떤 처치의 예방접종 효과, 즉 항체생성율이 비감염율 그리고 예방접종율과 같은 관련비율들에 어떻게 영향을 받는가를 알아보기 위한 통계적 분석모형으로 연속적 종속모형을 제시하고, 모형내 미지모수들을 추정하기 위한 방법을 논의하고 있다.

1. 서 론

임상의학 및 전염병학 분야에서 수집되는 자료들은 대부분 범주형 자료들로 분류된다. 이들 범주형 자료들은 연구유형에 따라 조사(survey), 실험연구(experimental study), 그리고 관측연구(observational study)로 부터 여러가지 자료수집 방법들을 통하여 수집된다. 수집된 자료의 분석을 위하여 타당한 모형설정에 근거한 분석방법은 모형에 근거하지 않은 분석방법보다 효율적이고 체계적인 분석방법을 제공하여 준다. 따라서, 수집된 자료의 유형별 특성, 구조 및 성격에 따라 자료를 효과적으로 분석하기 위한 여러 모형들이 많은 문헌에서 제시되고 있다. 예를 들면, Anderson and Aitkin(1985) 은 조사설문지의 분석을 위한 면담자들의 변동을 다루었다. 이들은 이원 지분 조사계획(two-level nested survey design)으로 발생하는 분산성분들을 추정하기 위하여 로지트 모형(logit model)에 대한 최우추정법을 제시했다. Im and Gianola(1988)는 이원 지분 배열(two-way nested layout)로 부터의 이항자료에 대한 혼합모형에서의 최우추정치를 계산하기 위한 심플렉스 방법(simplex method)을 논의했다. Cox and Snell(1989)은 반응에 있어서 계층적 구조(hierarchical structure)를 갖는 다가자료(polytomous data)의 분석을 위하여 조건부 지분 확률변수의 이용에 관하여 논의했다. 또한, McCullagh and Nelder(1989)는 지분 구조(nested structure)를 갖는 다가자료(polytomous data)에 대한 가능한 모형들을 제시하고 있다. 여기서 다가자료란 하나의 반응값이 여러 가능한 범주중 한범주에만 분류되는 이산자료를 의미한다. 최재성(1996)은 조건부 지분 확률변수(conditional nested random variable)를 내포하고 있는 모형설정 및 모형내 모수들의 추론방법을 구체적으로 논의하고 있다. 확률효과를 갖는 이가자료를 분석하기 위하여 Conaway(1990)는 각 개체에 대한 반복측정치들이 독립임을 의미하는 국부적 독립모형과 반응들간의 추가종속성을 반영하기 위한 종속모수들을 포함하는

1) 본 연구는 1995년도 계명대학교 비사연구기금으로 이루어졌음.
2) (704-701) 대구광역시 달서구 신당동 계명대학교 자연과학대학 통계학과 부교수.

종속모형들을 다루었다.

인체에 관한 질병을 감염성 여부에 따라 감염성 질병과 비감염성 질병의 두 부류로 분류할 때, 이들 질병과 관련된 처치들의 효과에 대한 분석방법들은 질병의 감염성 여부를 고려하지 않고 주로 질병에 감염된 개체들만을 대상으로 그 질병을 치료하기 위한 처치들의 효과, 또는 질병에 감염되지 않은 개체들만을 대상으로 그 질병을 예방하기 위한 처치들의 효과를 분석하고자 한다. 관심질병이 감염성 질병일 때, 이를 예방하기 위한 개발약품의 처치효과는 질병발생 집단내 감염된 개체들의 수에 영향을 받을 수 있음을 최재성(1996)의 논문에서 보여진다.

본 논문은 관심질병이 감염성 질병이고 여러지역에서 다발적으로 발생하며, 또한 질병발생집단에서 영구면역을 갖는 개체들이 존재할때, 그 질병으로 부터의 감염을 예방하기 위하여 개발된 한 처치의 효과를 알아보기 위한 모형설정 및 추정방법을 논의하고자 한다. 고려되고 있는 감염성 질병에 대한 개발약품의 효과를 알아보기 위한 모형설정을 위하여 다음과 같은 실험상황을 가정하여 보자.

감염성의 조사질병은 주로 소아에서 발생하는 수두(Chickenpox, Varicella)라 하자. 수두(문희주 외, 1988; 홍창의, 1994)는 거의 대부분이 소아에서 발생하며, 전염력이 매우 강하고, 전신적인 발진을 동반하며, 감수성이 있는 집단에서는 급속한 유행을 일으키는 질환이다. 수두의 전파는 환자의 타액에 의한 비말 감염(droplet infection)이나, 직접 접촉을 통해 유행성으로 퍼진다. 호발연령은 5-9세이고, 계절적으로 늦겨울과 초봄에 호발하며, 한번의 현성감염 후는 일생동안 면역이 지속되는 것으로 알려져 있다. 수두를 예방하기 위한 약품을 제일제당에서 개발판매 한다고 하자. 이 약품의 예방효과를 알아보기 위하여, 인구 오만이상의 지역에서 유치원에 다니고 있는 미취학 아동들을 대상으로 그 효과를 조사한다고 가정하자. 예방접종할 어린이들을 추출하기 위하여 먼저 인구 오만 이상의 지역들의 모집단으로 부터 몇개 지역을 임의로 선정하고, 그다음 추출된 각 지역내에서 임의로 일부 유치원들을 추출한다. 추출된 유치원에서 유치원생들을 대상으로 먼저 수두에 대한 감염여부를 검사하여 조사시점에서 수두에 감염된 어린이들과 수두에 감염되지 않은 어린이들로 분류한 후 다시 감염되지 않은 어린이들을 대상으로 예방접종이 필요한 어린이들과 예방접종이 필요하지 않은 어린이들로 재분류한다. 세 번째로 예방접종이 필요한 감염되지 않은 유치원생들에 한하여 개발된 약품의 예방접종을 실시한다. 일정기간 이후 예방접종에 의하여 항체가 생긴 어린이들을 조사함으로써 표본으로 선정된 유치원내 어린이들의 항체생성비율을 추정할 수있다. 이 경우에, 실험구조로 부터 두 가지 추가적인 변동요인들이 발생하게 된다. 첫번 째 변동요인은 인구 오만이상의 지역집단에서 일부지역을 표본으로 추출함으로써, 지역간의 관심비율들의 변동을 예상할 수 있고, 두번 째 변동요인은 선정된 지역내에서 유치원들을 추출할 때 유치원간의 관심비율들의 변동을 생각할 수 있다. 따라서, 유치원생들 간의 개별적 차이, 또는 실험단위들 간의 차이 이외에도, 관측비율들은 지역간의 변동과 유치원간의 변동을 갖게된다. 이와같이 한 처치, 즉 개발된 약품의 예방접종, 또는 처치들의 실험단위들을 위에서 논의된 집락추출법에 의하여 얻게될 때 이원 지분계획법(two-way nested design)을 이용할 수 있다.

본 연구는 예방접종에 의한 항체생성율과 이와 관련된 확률들에 관심을 두고, 관계되는 확률들이 각기 고려중의 모집단에서 조건부확률로 정의되며, 예방접종에 의한 항체생성율은 예방접종한 개체들의 집단에서 예방접종에 의해 항체를 갖는 개체의 조건부확률로 정의될 수 있음을 보여준다. 즉, 항체생성율 및 관련확률들은 조건부 지분 확률변수(conditional nested random

variable)의 확률로써 표현될 수 있다. 본 논문은 개발약품의 처치효과를 평가하기 위하여 관측조사로 부터 수집된 자료에 조건부 지분 확률변수의 개념을 이용한 연속적인 종속모형을 설정하는 방법을 기술하고 있다.

타당한 모형설정을 위하여 조사대상 모집단에 대해 다음 가정들이 필요하다. 인구 오만이상의 거주지역에 살고있는 유치원생들의 모집단으로 부터 한 개체의 반응이 다음과 같이 가능한 네 범주중 하나로 분류된다고 하자.

A_1 은 조사시점에서 관측된 개체가 감염되었을 때의 범주이고,

A_2 는 조사시점에서 관측된 개체가 면역이 되어 있을 때의 범주이며,

A_3 는 조사시점에서 관측된 개체가 감염되지 않고, 예방접종후 항체가 생기지 않았을 때의 범주이고,

A_4 는 조사시점에서 관측된 개체가 감염되지 않고, 예방접종후 항체가 생겼을 때의 범주이다.

$A_2 \bigcup A_3 \bigcup A_4$ 는 감염되지 않은 개체들의 집단을 나타내고, $A_3 \bigcup A_4$ 는 예방접종한 개체들의 집단을 나타내므로 반응들은 지분 구조를 갖게 되고, 이러한 반응에 있어서의 지분 구조(nested structure)는 조건부 지분 확률변수를 정의함으로써 타당한 모형전개에 이용될 수 있다.

실험단위들을 집락추출법으로 얻을 때, A_1, A_2, A_3, 그리고 A_4 의 각 범주에 속할 유치원생들의 반응확률은 유치원간의 변동이 있게 된다. 즉, 네 범주에 속할 개체들의 확률은 유치원마다 다를 수 있다. 이러한 현상은 관심확률이 유치원간의 변동을 허용하는 모형으로 설명할 수 있다. 먼저, 유치원내에서는 각 개체들이 이들 네 범주에 속할 확률은 일정하나, 유치원간의 변동은 가능하다고 가정한다. 각 범주에 속할 확률들이 집락추출법에 따른 집락간에 변동하기 때문에, 두 가지 변동요인, 즉, 유치원 과 지역, 에 따른 초과변동(over-dispresion)이 발생할 수 있다.

비율에 영향을 미치는 몇 가지 변동요인들이 있을 때, 혼합모형 또는 확률모형의 근거하에 분산성분들을 추정할 수 있다. 혼합모형내 확률효과들의 분산성분들의 추정은 가우시안 구적점(Gaussian Quadruture points)을 이용한 심플렉스 알고리즘을 이용할 수있다. 수치적분을 위하여 가우스-허미트(Gauss-Hermite)공식을 이용할 때, $\int f(u)\phi(u)du$ 의 적분에 대한 M-point 가우시안 구적(Gaussian quadrature)은 다음과 같이 가중합으로 근사치가 구해진다.

$$\sum_{i=1}^{M} w_i f(x_i)$$

단, x_i 는 가우시안 구적점 이고 w_i 는 Abramowitz and Stegun(1972)에 의해 기술된 관련 비중들 이다.

2. 모형

표본추출방법을 고려한 이원 지분계획법(two-way nested design)으로 부터 자료를 수집한다고 가정하자. i 를 지역, {1, 2, ... , i, ... , I}, 에 대한 지수라 두고, j 를 지역내 유치원, {1, 2,

... , j, ..., J_I}, 그리고 k 를 유치원내 어린이, {1, 2, ... , n_{ij}, ... , n_{IJ_I}}, 에 대한 지수라 두자. 지역 i 를 임의로 선정한 후, 유치원(i, j)가 지역 i 내에서 임의로 추출되고 유치원(i, j)내 n_{ij} 명의 어린이들로 부터 자료를 수집한다. 관측조사에서 각 유치원내 수두에 대한 예방이 필요한 어린이들에게 행해진 예방접종을 처치라 하자. 유치원(i, j)에 대해 조건부 지분 확률변수를 다음과 같이 정의한다. 단, 감염여부는 조사시점에서 행해진다.

$$U_{ijk}=\begin{cases}1 & \text{개체(i, j, k)가 감염되지 않았을 때}\\ 0 & \text{개체(i, j, k)가 감염되었을 때}\end{cases}$$

$$V_{ijk}=\begin{cases}1 & \text{개체(i, j, k)가 면역이 되지 않았을 때}\\ 0 & \text{개체(i, j, k)가 면역이 되었을 때}\end{cases}$$

$$W_{ijk}=\begin{cases}1 & \text{개체(i, j, k)가 예방접종후 항체가 생겼을 때}\\ 0 & \text{개체(i, j, k)가 예방접종후 항체가 생기지 않았을 때}\end{cases}$$

단, U_{ijk} 는 k번 째 개체의 감염상태를 나타내는 이가 확률변수(binary random variable)이다. V_{ijk} 는 U_{ijk}=1이 주어졌을 때 조건부 지분 확률변수로 정의되고, W_{ijk} 는 U_{ijk}=1 이고 V_{ijk}=1이 주어졌을 때 조건부 지분 확률변수로 정의된다. 그러나 U_{ijk}=0 일 때, 확률변수 V_{ijk} 와 W_{ijk} 는 정의되지 않는다. 또한 U_{ijk}=1 이고 V_{ijk}=0 일 때, W_{ijk} 는 정의되지 않는다. 따라서, 세 확률변수에 의해 생성된 표본공간, S, 는 다음과 같다.

$$S=\{(0,\ \varphi,\ \varphi),\ (1,\ 0,\ \varphi),\ (1,\ 1,\ 0),\ (1,\ 1,\ 1)\}.$$

단, 표본점 (0, φ, φ)의 의미는 U_{ijk}=0 일 때, 확률변수 V_{ijk} 와 W_{ijk} 는 정의되지 않음을 나타내고, 표본점 (1, 0, φ)의 의미는 U_{ijk}=1 이고 V_{ijk}=0 일 때, W_{ijk} 는 정의되지 않음을 나타낸다. 조건부 지분 확률변수들이 정의되지 않는 두 표본점,
(0, φ, φ) 와 (1, 0, φ), 들을 표기의 편의상 각기 (0, 0, 0) 와 (1, 0, 0)로 표기해도 무방하다. 이때 주어진 표본공간을 S_1 이라 두자. 즉,

S_1 = { (u_{ijk}, v_{ijk}, w_{ijk}); (0, 0, 0), (1, 0, 0), (1, 1, 0), (1, 1, 1) } 이다.

S_1 의 부분집합, S_2 는 U_{ijk}=1 이 주어졌을 때 V_{ijk} 의 값들에 대한 조건부 표본공간이다. 즉,

S_2 = { (u_{ijk}, v_{ijk}, w_{ijk}); (1, 0, 0), (1, 1, 0), (1, 1, 1)} 이다.

S_1 의 부분집합, S_3 는 U_{ijk}=1 이고 V_{ijk}=1 일 때 W_{ijk} 의 값들에 대한 조건부 표본공간이다. 즉,

S_3 = { (u_{ijk}, v_{ijk}, w_{ijk}); (1, 1, 0), (1, 1, 1)} 이다.

관심집단에 대한 관측조사로 부터의 네 범주, A_1, A_2, A_3, 그리고 A_4 는 표본공간을 네개의 상호배반인 사상들로 분할하는 S_1 의 부분집합들로 기술된다. 즉,

A_1 = { (u_{ijk}, v_{ijk}, w_{ijk}); (0, 0, 0) },
A_2 = { (u_{ijk}, v_{ijk}, w_{ijk}); (1, 0, 0) },
A_3 = { (u_{ijk}, v_{ijk}, w_{ijk}); (1, 1, 0) },

$$A_4 = \{ (u_{ijk}, v_{ijk}, w_{ijk}); (1, 1, 1) \}.$$

각 유치원에서의 반응들은 확률변수 U_{ijk}, V_{ijk}, 그리고 W_{ijk} 에 의해 표시된 네 범주로 분할된다. 지역 i 에서 j 번째 유치원과 관계된 사상들의 확률을 다음과 같이 두자.

$$\pi_{aij} = P(A_2 \cup A_3 \cup A_4),\ \pi_{bij} = P(A_3 \cup A_4),\ \text{그리고}\ \pi_{tij} = P(A_4).$$

따라서, π_{aij}는 지역 i 에서 j 번째 유치원내 감염되지 않은 개체들의 비율을 나타내며, π_{bij}는 지역 i 에서 j 번째 유치원내 감염되지 않았으나 예방접종한 개체들의 비율이고, π_{tij}는 지역 i 에서 j 번째 유치원내 감염되지 않았으나 예방접종후 항체가 생긴 개체들의 비율을 나타낸다.

다른 근원사상들에 대한 확률들은 이들 확률로 부터 다음과 같이 구해진다.

$$P(A_1) = 1 - \pi_{aij},\ P(A_2) = \pi_{aij} - \pi_{bij},\ \text{이고}\ P(A_3) = \pi_{bij} - \pi_{tij}\ \text{이다.}$$

항체생성율 및 관련비율들은 조사시점에서 관심질병에 감염되지 않은 개체들의 집단만을 생각한 비율이기 때문에, 이들 비율은 질병발생집단에서 조건부확률로 정의된다. 조건부확률을 전개하기 전에 표현의 단순화를 위하여, U_{ijk}, V_{ijk}, 와 W_{ijk} 에 해당하는 변수들을 각각 다음과 같이 정의한다.

$$I_{(1)}(U_{ijk}) = \begin{cases} 1 & U_{ijk} = 1 \text{ 이면,} \\ 0 & \text{그렇지 않으면} \end{cases}$$

$$I_{(1)}(V_{ijk}) = \begin{cases} 1 & U_{ijk} = 1 \text{ 이고 } V_{ijk} = 1 \text{ 이면,} \\ 0 & \text{그렇지 않으면} \end{cases}$$

그리고

$$I_{(1)}(W_{ijk}) = \begin{cases} 1 & U_{ijk} = 1,\ V_{ijk} = 1,\ \text{이고 } W_{ijk} = 1 \text{ 이면,} \\ 0 & \text{그렇지 않으면} \end{cases}$$

라 둔다.

그다음, 관심사상들에 관한 비율들의 관련성을 나타내기 위하여 이들 확률을 다음과 같이 정의한다.

$$p_{aij} = \pi_{aij},\ p_{bij} = \pi_{bij} / \pi_{aij},\ \text{이고}\ p_{tij} = \pi_{tij} / \pi_{bij}.$$

위의 정의로 부터 p_{aij} 는 유치원(i, j)에서 추출된 한 개체가 감염되지 않을 확률이고, p_{bij}는 유치원(i, j)에서 추출된 개체가 조사시점에서 감염되지 않았다면, 예방접종을 할 조건부 확률이고, p_{tij} 는 유치원(i, j)에서 추출된 개체가 예방접종을 하였다면, 그 개체가 예방접종후 항체를 가질 조건부확률 이다.

일반적으로, 이가자료 또는 비율로 표시된 집단화한 이가자료에 대한 모형은 적절한 연결함수를 이용하여 관련 확률의 변환값에 가법적인 선형모형으로 표현된다. 연결함수는 (0, 1)의 구간을 $(-\infty, \infty)$의 구간으로 대응시키는 미분가능한 단조함수로 정의한다.

이원 지분계획법으로 인하여 관측범주들에 대한 확률의 변동을 허용하고 있기 때문에 조건부

지분 변수에 대한 반응확률 또한 지분(nested) 계획의 확률효과로 인한 영향을 나타낸다. 확률변수 U_{ijk}, V_{ijk}, 와 W_{ijk} 의 반응들은 이가(binary)이고, 이들 세 확률변수들로 정의된 비율에 영향을 미치는 두 가지 확률효과가 존재한다. 이들은 지역 i 의 효과, L_i , 와 지역 i 내 j 번째 유치원의 효과, K_{ij}, 이다. 고정효과뿐만 아니라 확률효과의 함수로써 반응확률을 모형화하는 일반적인 방법은 적절히 선택한 연결함수, $g(\cdot)$, 로 일반화된 혼합모형을 이용하는 것이다. 적절히 선택된 연결함수는 고정효과와 확률효과의 가법적인 선형함수로 표현되는 변환된 확률의 예측값을 제공한다. 이가변수들의 지분구조(nested structure)로 부터 연속적인 모형을 전개할 수 있다.

전염성이 강한 질병, 즉, 유행성 수두에 대한 연속적인 모형은 다음 모형으로써 표현될 수 있다.

$$
\begin{aligned}
g[P(I_{(1)}(U_{ijk}) = 1 \mid h_{ij}, l_i)] &= g(p_{aij}) = \alpha_1 + \beta_1 n_{ij} + l_i + h_{ij}, \\
g[P(I_{(1)}(V_{ijk}) = 1 \mid h_{ij}, l_i)] &= g(p_{bij}) = \alpha_2 + \beta_2 y_{aij} + l_i + h_{ij}, \\
g[P(I_{(1)}(W_{ijk}) = 1 \mid h_{ij}, l_i)] &= g(p_{tij}) = \alpha_3 + \beta_3 y_{bij} + \beta_4(y_{aij} - y_{bij}) + \beta_5(n_{ij} - y_{aij}) \\
&\quad + li + hij,
\end{aligned}
\tag{1}
$$

단, α_1, α_2, 와 α_3 는 각 선형예측의 절편이고 β들은 회귀모수들 이다. n_{ij}는 유치원(i, j)내 어린이들의 수를 나타내고, y_{aij}는 조사시점에서 감염되지 않은 어린이들의 수 이고, y_{bij}는 유치원(i, j)내 예방접종한 어린이들의 수이고, y_{tij}는 유치원(i, j)내 예방접종후 항체가 생성된 어린이들의 수를 나타낸다.

유치원의 특성에 따른 많은 변수들이 모형에 포함될 수 있으나, 변수들의 선정과 갯수는 실험환경에 달려있다. 이 모형에서는 단지 감염개체들의 함수만을 고려한다.

확률효과와 고정효과간의 차이는 주로 표본추출법에 의해 결정된다. 따라서, h_{ij} 와 l_i 는 확률효과를 나타내고 α와 β들은 고정효과를 나타낸다. 확률효과들인 경우, $\{H_{ij}\}$ 는 평균이 0 이고 분산이 σ_h^2인 독립이고 동일한 정규분포를 따른다고 가정한다. $\{L_i\}$ 또한 독립이고 동일분포, $N(0, \sigma_l^2)$, 를 따르며 $\{H_{ij}\}$ 와 $\{L_i\}$ 는 독립이라 가정한다.

대다수의 적용에서 동일한 연결함수를 모든 반응변수들에 이용할 수 있으나, 반드시 동일한 연결함수를 이용할 필요는 없다. 연결함수를 선정한후 모형내 모수들을 추정한다.

3. 모수의 추정

관측조사에 대한 다항확률벡타, 추출된 유치원(i, j)에서 개체(i, j, k)에 대한 $(U_{ijk}, V_{ijk}, W_{ijk})$, 의 결합확률분포는

$$
f(u_{ijk}, v_{ijk}, w_{ijk} \mid \pi_{aij}, \pi_{bij}, \pi_{tij}) =
\begin{cases}
\pi_{tij}^{w_{ijk}} (\pi_{bij} - \pi_{tij})^{v_{ijk} - w_{ijk}} (\pi_{aij} - \pi_{bij})^{u_{ijk} - v_{ijk}} (1 - \pi_{aij})^{1 - u_{ijk}} \\
\qquad\qquad \text{for} (u_{ijk}, v_{ijk}, w_{ijk}) \in S_1 \\
0 \qquad\qquad \text{그렇지 않으면}
\end{cases}
$$

이다.

2절에서 각 범주에 속할 반응확률들, $P(A_1)$, $P(A_2)$, $P(A_3)$, 와 $P(A_4)$, 는 U_{ijk}, V_{ijk}, 와 W_{ijk}의 결합분포로 기술될 수 있음에 유의한다. 결합분포로 부터, 확률변수 U_{ijk}, V_{ijk} 와 W_{ijk} 의 주변 확률분포는 각기 성공확률 p_{aij}, p_{bij}, 와 p_{tij}인 베르누이분포를 따름을 알 수 있다.

관심비율들에 대한 정보는 추출된 유치원에서의 자료벡타로 부터 얻어지기 때문에 일부 확률변수들을 정의할 필요가 있다. 따라서, (i, j)번째 유치원에 대해 Y_{aij}, Y_{bij} 와 Y_{tij}를 다음과 같이 정의한다.

Y_{aij} = 조사시점에서 수두에 감염되지 않은 개체들의 수,
Y_{bij} = 조사시점에서 감염되지 않고 예방접종을 필요로 하는 개체들의 수,
Y_{tij} = 조사시점에서 감염되지 않고 예방접종후 항체가 생긴 개체들의 수,

라 두자. 표본으로 뽑혀진 유치원(i, j)에서 Y_{aij}, Y_{bij}, 와 Y_{tij}의 결합분포는 모수가 n_{ij}, π_{aij}, π_{bij}, 그리고 π_{tij}인 다항분포임을 확률변수 U_{ijk}, V_{ijk} 와 W_{ijk}의 정의로 부터 쉽게 입증된다. 다음 결과로 부터 다항분포를 따름을 알 수 있다.

결과1]: 유치원(i, j)에서 어린이들의 수, n_{ij}, 를 알 때, Y_{aij}, Y_{bij}, 와 Y_{tij}의 결합분포는

$$f_{ij}(y_{aij}, y_{bij}, y_{tij}|n_{ij}, \pi_{aij}, \pi_{bij}, \pi_{tij}) =$$

$$c_{ij}\ \pi_{tij}^{y_{tij}}\ (\pi_{bij}-\pi_{tij})^{y_{bij}-y_{tij}}\ (\pi_{aij}-\pi_{bij})^{y_{aij}-y_{bij}}\ (1-\pi_{aij})^{n_{ij}-y_{aij}}$$

단, $c_{ij} = n_{ij}!/[y_{tij}!(y_{bij}-y_{tij})!(y_{aij}-y_{bij})!(n_{ij}-y_{aij})!]$ 이다.

증명1]: 확률변수, Y_{aij}, 는 모수 n_{ij} 와 π_{aij}를 갖는 이항분포를 따른다. Y_{aij}의 확률함수를 $h_1(y_{aij})$ 라 두자. $Y_{aij}=y_{aij}$ 가 주어졌을 때, Y_{bij} 와 Y_{tij} 의 조건부 결합분포는

$$h_2(y_{tij}, y_{bij}|y_{aij}) = (y_{aij}!/[y_{tij}!(y_{bij}-y_{tij})!(y_{aij}-y_{bij})!])\ (\pi_{tij}/\pi_{aij})^{y_{tij}}$$

$$((\pi_{bij}-\pi_{tij})/\pi_{aij})^{y_{bij}-y_{tij}}\ (1-(\pi_{bij}/\pi_{aij}))^{y_{aij}-y_{bij}}$$

이다. 따라서, Y_{aij}, Y_{bij}, 와 Y_{tij}의 결합분포는

$f_{ij}(y_{tij}, y_{bij}, y_{aij}|n_{ij}, \pi_{tij}, \pi_{bij}, \pi_{aij}) = h_1(y_{aij})h_2(y_{tij}, y_{bij}|y_{aij})$ 이므로 위의 결과를 얻는다.

관련비율들을 세 변수, Y_{aij}, Y_{bij}, 와 Y_{tij}, 의 반응확률로써 정의하는데 관심을 두고 있기 때문에, 다음 결과는 무조건부확률을 조건부 지분 이가 변수(conditional nested binary variable)로 정의된 확률로 치환된 경우를 나타낸다.

결과2]: 모수로 n_{ij}, p_{aij}, p_{bij} 와 p_{tij}를 갖는 Y_{aij}, Y_{bij} 와 Y_{tij} 의 결합분포는

$$g_{ij}(y_{aij}, y_{bij}, y_{tij}|n_{ij}, p_{aij}, p_{bij}, p_{tij}) = c_{ij}\ p_{tij}^{y_{tij}}\ (1-p_{tij})^{y_{bij}-y_{tij}}\ p_{bij}^{y_{bij}}$$

$$(1-p_{bij})^{y_{aij}-y_{bij}}\ p_{aij}^{y_{aij}}\ (1-p_{aij})^{n_{ij}-y_{aij}} \qquad (2)$$

단, $c_{ij} = n_{ij}!/[y_{tij}!(y_{bij}-y_{tij})!(y_{aij}-y_{bij})!(n_{ij}-y_{aij})!]$ 이다.

증명2]: 위 결과는 π들의 항들로 주어진 p에 관한 정의를 이용함으로써 입증된다. 즉,

$$f_{ij}(y_{aij}, y_{bij}, y_{tij}|n_{ij}, \pi_{aij}, \pi_{bij}, \pi_{tij}) =$$

$$c_{ij}\ \pi_{tij}^{\,y_{tij}}\ (\pi_{bij}-\pi_{tij})^{y_{bij}-y_{tij}}\ (\pi_{aij}-\pi_{bij})^{y_{aij}-y_{bij}}\ (1-\pi_{aij})^{n_{ij}-y_{aij}}$$

로 부터 π들을 p들로 변환하기 위하여 각 π에 어떤 양을 곱하고 나누어 주면,

$$f_{ij}(y_{aij}, y_{bij}, y_{tij}|n_{ij}, \pi_{aij}, \pi_{bij}, \pi_{tij}) =$$

$$c_{ij}\ (\pi_{tij}/\pi_{bij})^{y_{tij}}\ \pi_{bij}^{\,y_{tij}}\ [(\pi_{bij}-\pi_{tij})/\pi_{bij}]^{y_{bij}-y_{tij}}\ \pi_{bij}^{\,y_{bij}-y_{tij}}$$

$$[(\pi_{aij}-\pi_{bij})/\pi_{aij}]^{y_{aij}-y_{bij}}\ \pi_{aij}^{\,y_{aij}-y_{bij}}\ (1-\pi_{aij})^{n_{ij}-y_{aij}}$$

이므로 위의 결과를 얻을 수 있다.

각 유치원(i, j)의 자료벡타 $(y_{aij}, y_{bij}, y_{tij})$에 대한 결합분포를 결과2]로 부터 알 수 있기 때문에 연속모형내 미지모수들을 최우법으로 추정할 수 있다.

$\mathbf{Y}_{ij}=(Y_{aij}, Y_{bij}, Y_{tij})$ 를 유치원(i, j)에 대한 확률벡타라 두자. H=h 와 L=l 이 주어졌을 때, $\mathbf{Y} = (\boldsymbol{Y}_{11}, \boldsymbol{Y}_{12}, \ldots, \boldsymbol{Y}_{IJ})$의 조건부분포는

$$f(\boldsymbol{y}: \boldsymbol{\theta}, \boldsymbol{h}, \boldsymbol{l}) = \prod_i \prod_j g_{ij}(y_{aij}, y_{bij}, y_{tij}|H_{ij}=h_{ij}, L_i=l_i, \boldsymbol{\theta})$$

이다. 단, $\boldsymbol{\theta}=(\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$는 모형(1)내 미지 모수벡타이고,

$\boldsymbol{H}=(H_{11}, H_{12}, \ldots, H_{IJ})$, 이고 $\boldsymbol{L}=(L_1, L_2, \ldots, L_I)$ 이다.

(H, L)의 결합밀도함수는

$$\mathbf{g}(\mathbf{h}, \mathbf{l}) = \prod_i [\phi_1(l_i) \prod_j \phi_2(h_{ij})]$$

로 주어진다. 단, $\phi_1(l_i)$는 평균이 0이고 분산이 $\sigma_L{}^2$인 정규분포를 따르고, $\phi_2(h_{ij})$는 평균이 0이고 분산이 $\sigma_h{}^2$인 정규분포를 따르며, l_i 와 h_{ij}는 모든 (i, j)에 대해 독립이라고 가정한다.

이때, (Y, H, L)의 무조건부 분포는

$$\begin{aligned} f(\boldsymbol{y}: \boldsymbol{\theta}, \boldsymbol{h}, \boldsymbol{l}) &= [\prod_i \prod_j g_{ij}(y_{aij}, y_{bij}, y_{tij}|H_{ij}=h_{ij}, L_i=l_i, \boldsymbol{\theta})] \\ &\quad \{\prod_i [\phi_1(l_i) \prod_j \phi_2(h_{ij})]\} \\ &= \prod_i \{\phi_1(l_i)[\prod_j g_{ij}(\boldsymbol{y}_{ij}|H_{ij}=h_{ij}, L_i=l_i, \boldsymbol{\theta})\phi_2(h_{ij})]\}. \end{aligned} \tag{3}$$

이다.

Y의 주변확률분포는

$$f_y(\boldsymbol{y},\ {\sigma_H}^2,\ {\sigma_L}^2) = \prod_i \Big\{ \int \phi_1(l_i)[\prod_j \int g_{ij}(\boldsymbol{y}_{ij}|H_{ij}=h_{ij},\ L_i=l_i,\ \boldsymbol{\theta}) \\ \phi_2(h_{ij})\, dh_{ij}]\, dl_i \Big\}.$$

이나, Y의 주변 로그우도는

$$MLLH = \log f_y(\boldsymbol{y}:\boldsymbol{\theta},\ {\sigma_H}^2,\ {\sigma_L}^2) \\ = \sum_i \log\Big\{ \int \phi_1(l_i)[\prod_j \int g_{ij}(\boldsymbol{y}_{ij}|H_{ij}=h_{ij},\ L_i=l_i,\ \boldsymbol{\theta}) \\ \phi_2(h_{ij})\, dh_{ij}]\, dl_i \Big\} \tag{4}$$

이다.

추정방정식들은 Y의 주변 로그우도, *MLLH*(marginal log-likelihood), 를 미지모수들에 관하여 미분함 으로써 구해지며, 이들 방정식들은 모수들 간에 비선형이다. 따라서, 최우추정치들은 Dempster et al(1977)의 EM 알고리즘, 또는 Nelder and Mead(1965)의 심플렉스 방법과 같은 반복적인 수치방법으로 구해진다.

4. 로지트-연결(logit-link) 모형

일반적으로, 로지스틱 모형은 주로 자료에서 계산된 대수 승산비의 해석을 위해 이용된다. 만일 로지스틱 모형이 한 변수의 선형함수이면, 반응확률은

$$p = P(\text{반응}) = e^{(\alpha+\beta x)}/[1+e^{(\alpha+\beta x)}]$$

가 되고, 로지스틱 연결함수는 $\log[p/(1-p)] = \alpha+\beta x$ 로 모형을 선형화 한다. 이 경우, $\log[p/(1-p)]$를 대수 승산이라 하고, 이용된 함수를 로지트 연결함수라 한다. 전염성이 강한 질병에 관한 관측조사로 부터의 관심비율들을 로지트 연결함수로 나타낼 때, 2절에서 모형(1)은

$$\text{logit}\, p_{aij} = \alpha_1 + \beta_1\, n_{ij} + l_i + h_{ij},$$
$$\text{logit}\, p_{bij} = \alpha_2 + \beta_2\, y_{aij} + l_i + h_{ij},$$
$$\text{logit}\, p_{tij} = \alpha_3 + \beta_3\, y_{bij} + \beta_4(y_{aij}-y_{bij}) + \beta_5(n_{ij}-y_{aij}) + l_i + h_{ij},$$

로 표현된다. 감염된 개체들의 수가 예방접종후 항체가 생긴 개체들에 영향을 미치는 가의 여부에 관계된 미지모수는 β_5 이다. 만일 $\beta_5=0$ 이면, 항체생성율은 감염개체들의 수에 영향을 받지 않음을 의미한다.

로지트 모형내 미지모수들에 관한 추론은 확률효과를 나타내는 두 확률변수, L_i 와 H_{ij}, 의 분산성분들을 모형내 포함하기 위하여 $A_i=L_i/\sigma_L$ 와 $B_{ij}=H_{ij}/\sigma_H$ 로 정의된 두 확률변수 A_i 와 B_{ij} 를 도입한 아래 연속모형에서 행해질 수 있다.

$$\text{logit}\, p_{aij} = \alpha_1 + \beta_1\, n_{ij} + \sigma_L a_i + \sigma_h b_{ij},$$

$$\text{logit}\, p_{bij} = \alpha_2 + \beta_2\, y_{aij} + \sigma_L a_i + \sigma_h b_{ij}, \qquad (5)$$
$$\text{logit}\, p_{tij} = \alpha_3 + \beta_3\, y_{bij} + \beta_4(y_{aij}-y_{bij}) + \beta_5(n_{ij}-y_{aij}) + \sigma_L a_i + \sigma_h b_{ij}.$$

두 확률변수 A_i 와 B_{ij}는 각기 표준정규분포를 따르기 때문에, Y의 주변 로그우도는

$$l(\theta, \sigma_H, \sigma_L | y) = \sum_i \log\Big\{ \int \phi(a_i)[\prod_j \int c_{ij}\, p_{tij}^{\,y_{tij}}\, (1-p_{tij})^{\,y_{bij}-y_{tij}}$$
$$p_{bij}^{\,y_{bij}}\, (1-p_{bij})^{\,y_{aij}-y_{bij}}\, p_{aij}^{\,y_{aij}}\, (1-p_{aij})^{\,n_{ij}-y_{aij}} \qquad (6)$$
$$\phi(b_{ij})\, db_{ij}]\, da_i \Big\}$$

단, $c_{ij} = n_{ij}!/[y_{tij}!(y_{bij}-y_{tij})!(y_{aij}-y_{bij})!(n_{ij}-y_{aij})!]$이고, $\phi(\cdot)$는 표준 정규 밀도 함수이다.

위 함수를 최대로 하는 방법은 일반적으로 EM 알고리즘을 이용할 수 있으나, 이 방법은 모수 추정치들의 표준오차를 나타내지 않고, 수렴이 늦기 때문에 Im and Gianola(1988)은 다른방법 으로써 심플렉스 방법(Griffiths and Hill, 1985)을 제시하고 있다.

5. 로지트-연결 모형에 대한 예

다음도표는 생성자료를 표시한다. 이 절의 목적은 표5.1의 생성자료에 4절의 식(5)를 적합시켰을 때에 본문에서 논의된 모수들의 추정에 관한 이론 및 방법을 이용하여 구할 수 있음을 구체적으로 설명하고 있다. 또한 제시된 모형내 한 관심모수의 검정방법을 보여주고 있다. 지역 i, i=1, 2, 3, 와 유치원 j, j=1, 2, 3, 4, 에 대한 (i, j) 번째 유치원에서 개체들의 수(n_{ij}), 조사시점에서 감염되지 않은 개체들의 수(y_{aij}), 예방접종한 개체들의 수 (y_{bij}), 그리고 예방접종후 항체가 생긴 개체들의 수(y_{tij})를 나타낸다.

표5.1 감염성 수두에 대한 자료

지역	유치원	n_{ij}	y_{aij}	y_{bij}	y_{tij}
1	1	57	32	13	6
	2	50	30	11	6
	3	50	21	7	4
	4	53	27	10	6
2	1	55	20	5	2
	2	59	30	7	0
	3	52	12	1	1
	4	51	17	1	0
3	1	54	40	27	23
	2	58	46	34	29
	3	52	45	31	27
	4	51	43	32	43

4절에서 논의된 로지트 모형을 표 5.1의 자료에 적합시킬 때, 자료베타 **Y**의 주변 로그우도는

$$L(\boldsymbol{\theta}, \sigma_H, \sigma_L|\boldsymbol{y}) = \sum_i \log\Big\{\int \phi(a_i)[\prod_j \int g_{ij}^*(y_{aij}, y_{bij}, y_{tij}|A_i = a_i, B_{ij} = b_{ij})\phi(b_{ij})\, db_{ij}]\, da_i\Big\},$$

단, g_{ij}^* 는 $A_i = a_i$ 와 $B_{ij} = b_{ij}$가 주어졌을 때, Y_{aij}, Y_{bij} 와 Y_{tij}의 다항분포이다.

최우추정치들은 주변 로그우도 함수에 근거를 두고 있기 때문에, 최우추정치를 얻기 위한 Nelder and Mead(1965)의 심플렉스 방법에서 근사적인 음의 로그우도 함수를 이용한다.

구적점(quadrature point)의 개수를 M으로 두자. Brillinger and Preisle(1983)에 따르면, M이 여덟 개 이상 이면 최우추정치들은 크게 변하지 않으므로, 표 5.2는 표 5.1의 생성자료에 대한 분석을 위해 M=8을 이용했을 때 로지트 모형내 모수들의 최우추정치와 표준오차는 다음과 같이 구해진다. 괄호안은 표준오차를 나타낸다.

표 5.2 M=8일 때 로지트 모형에 대한 최우추정치 와 표준오차

가설	모수	최우추정치(표준오차)
H_a	α_1	-3.3858772(1.1103115)
	β_1	0.0596331(0.0283086)
	σ_L	0.9167590(0.0966060)
	σ_H	0.0000162(0.1185030)
	α_2	-1.7649947(0.4800835)
	β_2	0.0273643(0.0143813)
	α_3	1.2409443(3.3650138)
	β_3	0.0010980(0.0623351)
	β_4	-0.1066975(0.0932273)
	β_5	0.0083649(0.0657050)
	-MML	78.3468414
H_0: $\beta_5 = 0$	α_1	-3.3842355(1.5255177)
	β_1	0.0596095(0.0283048)
	σ_L	0.9159682(0.0964042)
	σ_H	0.0000002(0.1164556)
	α_2	-1.7664127(0.4799839)
	β_2	0.0274252(0.0143742)
	α_3	1.6077726(1.7430070)
	β_3	-0.0061336(0.0255732)
	β_4	-0.1111001(0.0868760)
	-MML	78.3549835

표 5.2로 부터 -MML은 최우추정치에서 계산된 주변 로그우도 함수의 음의 값을 나타낸다. 대립가설 H_a에 대해 귀무가설 H_0를 검정하기 위한 우도비 검정통계량은 $-2[MML(H_0)-MML(H_a)]$이다.

H_0를 검정하기 위한 검정통계량의 관측값은 $-2[MML(H_0)-MML(H_a)]=0.0162842$ 이고 근사적인 χ^2분포로 부터의 기각값은 $\chi^2_{(0.05,1)}$ 은 3.84 이기 때문에 H_0를 기각하지 않는다. 즉, H_0의 검정결과는 자료분석을 위한 모형에서 감염된 개체들의 수가 항체생성율에 영향을 미치지 않는 변수임을 의미하고 있다.

6. 결론

감염성 질병은 병후면역 상태로 분류할 때, 한 번의 현성감염후 면역이 일생동안 지속하는 질병과 잠시동안 면역을 나타내는 질병의 두 부류로 나눌 수 있다. 단기면역의 한 감염성 질병을 예로 든 최재성(1996)의 논문과는 달리 본 연구는 한 번의 현성감염후 일생동안 면역이 지속되는 질병중 하나인 수두에 대한 관측자료를 분석하기 위한 모형설정 과정및 모형내 미지모수들의 추정방법을 논의했다. 단기간의 면역을 갖는 감염성 질병의 경우는 질병발생 시기마다 미감염 개체들에 대해 예방접종을 실시할 수 있으나, 수두의 경우는 이 질병의 특성상 질병발생 집단내 예방접종이 필요하지 않은 이미 면역된 개체들의 집단이 존재함을 관측할 수 있다. 따라서, 단기면역의 감염성 질병에 대한 예방접종의 효과를 알아보기 위한 관측자료와는 다른 유형의 범주형 관측자료를 얻게된다. 이들 관측범주간의 지분구조는 단기간의 감염성 질병에 대한 관측범주들 간의 이가 지분구조와는 다름을 본문에서 제시하고 있다. 그러므로 자료를 분석하기 위한 모형도 새로운 지분구조를 토대로 모형설정이 이루어졌고, 모형내 포함된 미지모수들의 최우추정치를 구하기 위한 다항분포도 단기간의 면역을 갖는 질병의 경우와는 다른 다항분포를 이용해야 함을 제시하고 있다. 이는 또한 최우추정치 및 표준오차들을 구하기 위한 심플렉스 방법도 적절히 변형되었음을 의미한다. 결론적으로, 감염성 질병에 관한 한 예방백신의 관측자료를 분석할 때, 질병의 특성상 관측범주간의 지분구조가 서로 다른 경우에 서로 다른 모형토대 하에서 분석되어야 함을 알 수 있다.

참고문헌

[1] 문희주외 5인 (1988). 「면역혈청학」, 대학서림, 서울.

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

[3] 홍창의 (1994). 「소아과학」, 대한교과서, 서울.

[4] Abaramowitz, M. and Stegun, I. (1972). Handbook of mathematical functions, pp. 924., Dover Publications, New York.

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

[6] Brillinger, D. R. and Preisle, M. K. (1983). Maximum likelihood estimation in a latent variable problems. In studies in Economics, Time Series and Multivariate Statistics(eds S. Karlin, T. amemiya and L. A. Goodman), pp. 31-65, Academic Press, New York.

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

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

[9] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm(with discussion), *Journal of the Royal Statistical Society, Ser. B*, Vol. 39, 1-38.

[10] Griffiths, P. and Hill, I. D. (1985). Appied Statistics Algorithm, John Wiley & sons, New York.

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

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

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

A Generalized Linear Model For Vaccination Data On Chickenpox[3)]

Jaesung Choi[4)]

Abstract

This paper suggests a sequence of dependence models as a statistical analysis model for vaccination data on chickenpox and discusses a method for evaluating maximum likelihood estimates of unknown parameters in the suggested model.

3) The present research has been conducted by the Bisa Research Grant of Keimyung University in 1995.

4) Associate Professor, Department of Statistics, Keimyung University, 1000 Sindang-dong, Dalseogu, Taegu 704-701, Korea.