응용통계연구 제 9권 1호
1996년 3월 pp. 1-15.

질병의 범주적 자료에 대한 통계적 분석모형[1)]

최 재 성[2)]

요 약

본 논문은 질병발생집단의 감염율이 질병발생집단내 감염되지 않은 개체들에 대한 어떤 처치효과가 감염율에 어떻게 영향을 받는가를 알아보기 위한 통계적 분석모형으로 연속적 분석모형을 제시하고, 모형내 미지모수들을 추정하기 위한 방법을 논의하고 있다.

1. 서 론

질병을 다루는 의학, 수의학 및 Bioassay 분야에서 수집되는 자료들은 대부분 범주형 자료들로 분류된다. 질병과 관련된 범주형 자료들을 분석하기 위한 여러가지 통계적 분석방법들의 목적은 주로 질병에 감염된 개체들을 치유하기 위한 처치(약품 또는 수술)들의 치유율 또는 처치들의 효과를 알아보는 데 있다. 한편 면역학 또는 예방의학의 관점에서 본다면, 질병에 감염되지 않은 개체들을 조사질병으로 부터의 감염을 예방하기 위한 처치(예방접종, 또는 면역법)들에 대한 개체들의 항체생성률 또는 그 효과들을 추정하는 데 관심을 가질 수 있다. 이들 처치들의 효과를 알아보기 위하여 각기 질병에 감염된 개체들의 집단, 또는 질병에 감염되지 않은 개체들의 집단에서만 처치들을 비교하는 경우에 잘못된 추론을 할 수 있다. 왜냐하면, 이들 처치들의 효과가 질병발생집단의 감염율에 어떻게 영향을 받고 있는 가를 조사자는 알 수 없기 때문이다.

어떤 특정질병이 여러지역에서 발생할 경우 지역마다 그 질병에 감염될 감염율은 같지 않을 수 있음을 예상할 수 있다. 또한 감염율이 다른 지역에서 감염된 개체들의 감염정도가 지역간에 차이가 있을 때, 비교대상이 되는 처치들의 효과는 이들 감염율과 감염정도에 영향을 받게 된다. 따라서, 본 연구의 목적은 관심질병이 여러지역에서 다발적으로 발생할 때 그 질병에 대한 한 특정처치의 효과가 감염율에 영향을 받는 가를 알아보기 위한 모형설정과 함께 그 효과를 추정하는 방법을 논하고자 한다. 모형설정을 위한 예로써, 다음과 같은 실험상황을 생각하여 보자.

유행성독감을 예방하기 위한 한 약품을 녹십자에서 개발판매 한다고 하자. 일부 소아과 의사들은 환절기때 어린이들에게 이 약품의 예방접종을 적극 권장하고 다른 소아과 의사들은 이 약

---

1. 이 논문은 연구비중 일부를 1994년도 한국학술진흥재단의 공모과제 연구비에 의하여 지원받아 연구되었음.

2. (704-701) 대구광역시 달서구 신당동 계명대학교 자연과학대학 통계학과 부교수.

품의 효력을 자신할 수 없기 때문에 예방접종에 미온적이라 하자. 약품의 효과를 추론하기 위하여 인구 오만이상의 지역에서 유치원에 다니고 있는 미취학 아동들을 대상으로 그 효과를 알아보고자 한다고 가정하자. 여기서, 유행성독감은 두통, 근육통, 그리고 피로감의 갑작스러운 내습으로 묘사되는 특정한 바이러스 원인의 급성 호흡기 질환으로 정의된다. 유행성독감은 또한 환절기때 전국적으로 유행하는 호흡기질환으로 면역성이 약한 어린이들이나 노약자들에게는 치명적일 수도 있는 전염성이 강한 질병으로 알려져 있다. 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 유행성독감에 대한 감염여부를 검사한 다음 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 변동 과 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. 이러한 지분계획 구조(nested design structure)는 모형을 기술하기 위한 여러가지 방법들을 다룬 많은 다른 논문들에서 이용되고 있다. 예를들면, 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.

본 연구는 감염율과 항체생성율에 관심을 두고있기 때문에 항체생성율은 고려중의 모집단에서 감염되지 않은 개체의 조건부확률로 정의된다. 처치에 의한 항체생성개체들은 감염되지 않은 개체집단에서만 예방접종후 발생함으로 조건부 지분 확률변수(conditional nested random variable)의 확률로써 표현될 수 있다. 몇몇 저자들은 조건부 지분 확률변수를 이용했다. 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.

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. 확률효과를 갖는 이가자료를 분석하기 위하여 Conaway(1990)는 각 개체에 대한 반복측정치들이 독립임을 의미하는 국부적 독립모형과 반응들간의 추가종속성을 반영하기 위한 종속모수들을 포함하는 종속모형들을 다루었다. 그러나, 조건부 지분 확률변수(conditional nested random variable)를 내포하고 있는 모형에 관한 방법들은 문헌에서 찾아보기가 힘들다.

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 세 범주중 하나로 분류된다고 하자.

A_1 은 관측된 개체가 감염되었을 때의 범주이고,

A_2 는 관측된 개체가 감염되지 않았으나 예방접종후 항체가 생기지 않았을 때의 범주이며,

A_3 는 관측된 개체가 감염되지 않았으나 예방접종후 항체가 생겼을 때의 범주이다.

$A_2 \bigcup A_3$ 는 감염되지 않은 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, 그리고 A_3 의 각 범주에 속할 유치원생들의 반응확률은 유치원간의 변동이 있게된다. 즉, 세 범주에 속할 개체들의 확률은 유치원마다 다를 수 있다. 이러한 현상은 관심확률이 유치원간의 변동을 허용하는 모형으로 설명할 수 있다. 먼저, 유치원내에서는 각 개체들이 이들 세 범주에 속할 확률은 일정하나, 유치원간의 변동은 가능하다고 가정한다. 각 범주에 속할 확률들이 집락추출법에 따른 집락간에 변동하기 때문에, 두 가지 변동요인, 즉, 유치원과 지역에 따른 초과변동(over-dispresion)이 발생할 수 있다.

비율에 영향을 미치는 몇 가지 변동요인들이 있을 때, 혼합모형 또는 확률모형의 근거하에 분산성분들을 추정할 수 있다. 이항자료에 대한 혼합모형 과 확률모형을 초과 이항변동(extra-binomial variation)을 갖는 모형이라 부른다. 다양한 모형과 방법을 이용하여 초과 이항변동을 갖는 모형들에 대한 분산성분들을 추정하는 방법들을 일부 논문에서 찾아볼 수 있다.

Stiratelli, Laird and Ware(1984) 는 연속적인 이가반응들의 분석을 위한 일반적인 로지스틱 정규(logistic-normal) 모형을 제시하고, 확률모수들에 대하여 사후확률분포의 유형에 근거한 근사적인 방법을 제안했다. 확률효과와 분산성분들의 최우추정치는 EM 알고리즘 으로 구해졌고, 이 근사적인 방법은 각 개체에 대해 개별적인 로지스틱 회귀(logistic regression)를 이용하여 개체들의 반복된 이가건강치수들의 분석을 위한 Korn and Whittemore(1989)의 이단계법과 비교되었다.

Anderson and Aitkin(1985)은 로지트 모형에 대한 분산성분들을 추정하기 위하여 가우시안 구적점(Gaussian Quadruture points)을 이용한 EM 알고리즘을 실행했다. 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 관련비중들 이다.

2. 모형

지역내 속해있는 유치원들을 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 감염되지 않은 어린이들에게 행해진 예방접종을 처치라 하자. 유치원(i, j)에 대해 조건부 지분 확률변수를 다음과 같이 정의한다.

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

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

단, U_{ijk} 는 k번 째 개체의 감염상태를 나타내는 이가확률변수(binary random variable) 이다. V_{ijk} 는 U_{ijk}=1 이 주어졌을 때 조건부 지분 확률변수로 정의된다. 다시말하면, U_{ijk} 가 1의 값을 취할 때만 V_{ijk} 는 두 가지 가능한 값, 0 또는 1, 중 한 값을 취할 수 있다. U_{ijk}=0 일 때, 확률변수 V_{ijk} 는 정의되지 않는다. 두 확률변수에 의해 생성된 표본공간을 기술하기 위하여 표본점(0, Ø) 를 (0, 0)로 표기하자.

S_1 을 두 변수의 가능한 결과들의 표본공간이라 두자. 이때

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

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

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

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

$$A_1 = \{ (u_{ijk}, v_{ijk}); (0, 0) \},$$
$$A_2 = \{ (u_{ijk}, v_{ijk}); (1, 0) \},$$
$$A_3 = \{ (u_{ijk}, v_{ijk}); (1, 1) \}.$$

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

$$\pi_{ij} = P(A_2 \cup A_3),$$
$$\pi_{tij} = P(A_3).$$

다른 근원사상들에 대한 확률들은 이들 확률로 부터 다음과 같이 구해진다.

$P(A_1) = 1 - \pi_{ij}$, 이고

$P(A_2) = \pi_{ij} - \pi_{tij}$ 이다.

항체생성율은 조사질병에 감염되지 않은 개체들의 집단만을 생각한 비율이기 때문에, 감염율을 고려할 때, 이 비율은 질병발생집단에서 조건부확률로 정의된다. 조건부확률을 전개하기 전에 표현의 단순화를 위하여, U_{ijk} 와 V_{ijk} 에 해당하는 변수들을 각각 다음과 같이 정의한다.

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

로 두고,

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

라 둔다.

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

$$p_{ij} = \pi_{ij},$$
$$p_{tij} = \pi_{tij} / \pi_{ij}.$$

위의 정의로 부터 p_{ij} 는 유치원(i, j)에서 추출된 한 개체가 감염되지 않을 확률이고, p_{tij} 는 유치원(i, j)에서 추출된 개체가 감염되지 않았다면, 그 개체가 예방접종후 항체를 가질 조건부확률 이다.

일반적으로, 이가자료 또는 비율로 표시된 집단화한 이가자료에 대한 모형은 적절한 연결함수를 이용하여 관련확률의 변환값에 가법적인 선형모형으로 표현된다. 연결함수는 (0, 1)의구간을 $(-\infty, \infty)$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} 와 V_{ijk} 의 반응들은 이가(binary)이고, 이 두 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 방

법은 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 독감에 대한 연속적인 모형은 다음 모형으로써 표현될 수 있다.

$$g[P(I_{(1)}(U_{ijk}) = 1 \mid h_{ij}, l_i)] = g(p_{ij}) = \alpha_1 + \beta_1 n_{ij} + l_i + h_{ij},$$

$$g[P(I_{(1)}(V_{ijk}) = 1 \mid h_{ij}, l_i)] = g(p_{tij}) = \alpha_2 + \beta_2 y_{ij} + \beta_3(n_{ij}-y_{ij}) + l_i + h_{ij}, \qquad (1)$$

단, α_1, 과 α_2 는 각 선형예측의 절편이고 β들은 회귀모수들 이다. n_{ij}는 유치원(i, j)내 어린이들의 수를 나타내고, y_{ij}는 감염되지 않은 어린이들의 수 이고, y_{tij}는 유치원(i, j)내 예방접종 후 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}), 의 결합확률분포는

$$f(u_{ijk}, v_{ijk}|\pi_{ij}, \pi_{tij}) = \begin{cases} \pi_{tij}^{v_{ijk}}(\pi_{ij}-\pi_{tij})^{u_{ijk}-v_{ijk}}(1-\pi_{ij})^{1-u_{ijk}} & \text{for } (u_{ijk}, v_{ijk}) \in S_1 \\ 0 \quad \text{그렇지 않으면} \end{cases}$$

이다.

2절에서 각 범주에 속할 반응확률들, $P(A_1)$, $P(A_2)$, 와 $P(A_3)$, 은 U_{ijk}, 와 V_{ijk}의 결합분포로 기술될 수 있음에 유의한다. 결합분포로 부터, 확률변수 U_{ijk} 와 V_{ijk}는 각각 성공확률 p_{ij}, 와 p_{tij} 인 베르누이 분포를 따름을 알 수 있다. 예로써, U_{ijk}의 확률분포를 생각해 보자. U_{ijk}의 주변확률분포는 (U_{ijk}, V_{ijk}) 의 주어진 결합확률분포로 부터 U_{ijk}의 값들에 대한 결합확률을 합함으로써 구해진다. 이때,

$$f(u_{ijk})=\begin{cases} p_{ij} & u_{ijk}=1 \text{ 이면,} \\ 1-p_{ij} & u_{ijk}=0 \text{ 이면} \end{cases}$$

이다.

V_{ijk}는 조건부 지분 이가 확률변수(conditional nested binary random variable)로써 정의되기 때문에, 확률변수 U_{ijk}가 1의 값을 취할 때만이 V_{ijk}는 두가지 가능한 값, 0 또는 1, 중 한 값을 취한다. 따라서, V_{ijk}의 확률분포는 p_{tij}를 모수로 갖는 베르누이분포로 다음 확률함수를 갖는다.

$$f(v_{ijk})=\begin{cases} p_{tij} & v_{ijk}=1 \text{ 이면,} \\ 1-p_{tij} & v_{ijk}=0 \text{ 이면} \end{cases}$$

관심비율들에 대한 정보는 추출된 유치원에서의 자료벡타로 부터 얻어지기 때문에 일부 확률변수들을 정의할 필요가 있다.

(i, j)번째 유치원에 대해 Y_{ij}, 와 Y_{tij}를 다음과 같이 정의한다.

Y_{ij} = 감염되지 않은 개체들의 수,
Y_{tij} = 감염되지 않고 예방접종후 항체가 생긴 개체들의 수

라 두자.

표본으로 뽑혀진 유치원(i, j)에서 Y_{ij}, 와 Y_{tij}의 결합분포는 모수가 n_{ij}, π_{ij}, 그리고 π_{tij} 인 다항분포임을 확률변수 U_{ijk} 와 V_{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_{ij}, 와 Y_{tij}의 결합분포는

$$f_{ij}(y_{ij}, y_{tij}|n_{ij}, \pi_{ij}, \pi_{tij}) = c_{ij}\ \pi_{tij}^{y_{tij}}\ (\pi_{ij}-\pi_{tij})^{y_{ij}-y_{tij}}\ (1-\pi_{ij})^{n_{ij}-y_{ij}}$$

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

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

$$h_2(y_{tij}|y_{ij})=(y_{ij}!/[\ y_{tij}!(y_{ij}-y_{tij})!]\ )\ (\pi_{tij}/\pi_{ij})^{y_{tij}}\ (1-(\pi_{tij}/\pi_{ij}))^{y_{ij}-y_{tij}}$$

이다. 따라서, Y_{ij} 와 Y_{tij}의 결합분포는 $f_{ij}(y_{ij}, y_{tij}|n_{ij}, \pi_{ij}, \pi_{tij})=h_1(y_{ij})h_2(y_{tij}|y_{ij})$이므로 위의 결과를 얻는다.

관련비율들을 두 변수, Y_{ij} 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_{ij}, 와 p_{tij}를 갖는 Y_{ij} 와 Y_{tij} 의 결합분포는

$$g_{ij}(y_{ij}, y_{tij}|n_{ij}, p_{ij}, p_{tij}) = c_{ij}\ p_{tij}^{\,y_{tij}}\ (1-p_{tij})^{\,y_{ij}-y_{tij}}\ p_{ij}^{\,y_{ij}}\ (1-p_{ij})^{\,n_{ij}-y_{ij}} \quad (2)$$

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

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

$$f_{ij}(y_{ij}, y_{tij}|n_{ij}, \pi_{ij}, \pi_{tij}) = c_{ij}\ \pi_{tij}^{\,y_{tij}}\ (\pi_{ij}-\pi_{tij})^{\,y_{ij}-y_{tij}}\ (1-\pi_{ij})^{\,n_{ij}-y_{ij}}$$

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

$$f_{ij}(y_{ij}, y_{tij}|n_{ij}, \pi_{ij}, \pi_{tij}) = c_{ij}[\ (\pi_{tij}/\pi_{ij})^{\,y_{tij}}\ (\pi_{ij})^{\,y_{tij}}]\ [\ ((\pi_{ij}-\pi_{tij})/\pi_{ij})^{\,y_{ij}-y_{tij}}\ (\pi_{ij})^{\,y_{ij}-y_{tij}}]\ (1-\pi_{ij})^{\,n_{ij}-y_{ij}}$$

$$= c_{ij}[\ (\pi_{tij}/\pi_{ij})^{\,y_{tij}}\ (1-(\pi_{tij}/\pi_{ij}))^{\,y_{ij}-y_{tij}}]\ [\ \pi_{ij}^{\,y_{ij}}\ (1-\pi_{ij})^{\,n_{ij}-y_{ij}}]$$

이므로 위의 결과를 얻을 수 있다.

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

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

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

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

$\boldsymbol{H} = (H_{11}, H_{12}, \ldots, H_{IJ_I})$, 이고

$\boldsymbol{L} = (L_1, L_2, \ldots, L_I)$ 이다.

(H, L)의 결합밀도함수는

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

로 주어진다. 단, $\phi_1(l_i)$는 평균이 0이고 분산이 $\sigma_L{}^2$인 정규분포를 따르고, $\phi_2(h_{ij})$는 평

균이 0이고 분산이 σ_h^2인 정규분포를 따르며, l_i 와 h_{ij}는 모든 (i, j)에 대해 독립이라고 가정한다.

이때, (Y, H, L)의 무조건부 분포는

$$f(\boldsymbol{y}: \boldsymbol{\theta}, \boldsymbol{h}, \boldsymbol{l}) = [\prod_i \prod_j g_{ij}(y_{ij}, y_{tij}|H_{ij}=hij, L_i=l_i, \boldsymbol{\theta})] \{\prod_i [\phi_1(li)\prod_j \phi_2(h_{ij})]\}$$
$$= \prod_i \left\{\phi_1(l_i)[\prod_j g_{ij}(\boldsymbol{y}_{ij}|H_{ij}=h_{ij}, L_i=l_i, \boldsymbol{\theta})\phi_2(h_{ij})]\right\}. \quad (3)$$

이다.

Y의 주변확률분포는

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

이나, Y의 주변 로그우도는

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

이다.

추정방정식들은 Y의 주변 로그우도, $MLLH$, 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

$$\text{logit } p_{ij} = \alpha_1 + \beta_1 n_{ij} + l_i + h_{ij},$$

$$\text{logit } p_{tij} = \alpha_2 + \beta_2 y_{ij} + \beta_3(n_{ij}-y_{ij}) + l_i + h_{ij},$$

로 표현되고, 그 확률들은 모형으로 부터 다음과 같이 구해진다.

$$p_{ij} = \exp(\alpha_1 + \beta_1 n_{ij} + l_i + h_{ij})/[1+ \exp(\alpha_1 + \beta_1 n_{ij} + l_i + h_{ij})],$$

$$p_{tij} = \exp[\alpha_2 + \beta_2 y_{ij} + \beta_3(n_{ij}-y_{ij}) + l_i + h_{ij}]/\{1+\exp[\alpha_2 +\beta_2 y_{ij}+\beta_3(n_{ij}-y_{ij})+ l_i+ h_{ij}]\}.$$

로지트-연결 모형에 대한 로그 우도함수는 **Y**의 주변 로그우도에서 p_{ij} 와 p_{tij} 를 위의 식들로 대체하여 구한다. 최우추정치들은 수치방법을 이용하여 구한다.

감염된 개체들의 수가 예방접종후 항체가 생긴 개체들에 영향을 미치는 가의 여부에 관계된 미지모수는 β_3 이다. 만일 $\beta_3=0$ 이면, 항체생성율은 감염개체들의 수에 영향을 받지 않음을 의미한다.

로지트 모형내 미지모수들에 관한 추론은 확률효과를 나타내는 두 확률변수, L_i 와 H_{ij}, 의분산성분들을 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.

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

$$\text{logit } p_{tij} = \alpha_2 + \beta_2 y_{ij} + \beta_3(n_{ij}-y_{ij}) + \sigma_L a_i + \sigma_h b_{ij}. \tag{5}$$

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

$$l(\theta, \sigma_H, \sigma_L|\mathbf{y}) = \sum_i \log\Big\{\int \phi(a_{ij})[\prod_i \int c_{ij}\, p_{tij}^{\,y_{tij}}\,(1-p_{tij})^{\,y_{ij}-y_{tij}}\, p_{ij}^{\,y_{ij}} (1-p_{ij})^{\,n_{ij}-y_{ij}}\phi(b_{ij})\,db_{ij}]\;da_i\Big\} \tag{6}$$

단, $c_{ij} = n_{ij}!/[\,y_{tij}!(y_{ij}-y_{tij})!(n_{ij}-y_{ij})!]$ 이고, $\phi(\cdot)$ 는 표준 정규밀도 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 방법을 제시하고 있다.

5. 로지트-연결 모형에 대한 예

다음도표는 한 관측조사로 부터의 생성자료를 표시한다. 관측값들은 지역 i, i=1, 2, 3, 와 유치원 j, j=1, 2, 3, 4, 5, 6, 에 대한 (i, j) 번째 유치원에서 개체들의 수(n_{ij}), 감염되지 않은 개체들의 수(y_{ij}), 그리고 예방접종후 항체가 생긴 개체들의 수(y_{tij})를 나타낸다.

표5.1 유행성 독감 자료

지역	유치원	nij	yij	ytij
1	1	69	57	56
	2	62	49	48
	3	70	59	57
	4	69	56	55
	5	63	46	45
	6	68	57	56
2	1	66	52	51
	2	61	45	44
	3	60	41	14
	4	71	62	60
	5	61	42	23
	6	61	45	43
3	1	65	56	55
	2	61	43	37
	3	71	63	61
	4	63	51	51
	5	63	54	53
	6	66	53	52

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

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

단, g_{ij}^* 는 $A_i = a_i$ 와 $B_{ij} = b_{ij}$가 주어졌을 때, Y_{ij} 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	-3.886224(2.354841)
	β_1	0.081008(0.000560)
	σ_L	0.088537(0.015584)
	σ_H	0.153840(0.011171)
	α_2	14.411850(57.13300)
	β_2	-0.065892(0.009204)
	β_3	-0.576744(0.038194)
	-MML	67.361377
H_0: β_3 =0	α_1	-3.768746(0.484111)
	β_1	0.087504(0.000677)
	σ_L	0.987176(3.399873)
	σ_H	0.251513(0.007502)
	α_2	-7.606897(2.431471)
	β_2	0.219312(0.000917)
	-MML	75.469577

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)]=16.2164 이고 근사적인 χ^2분포로 부터의 기각값은 $\chi^2_{(0.05,1)}$ 은 3.84 이기 때문에 H_0를 채택하지 않는다.

6. 결론

질병발생 집단에서 어떤 처치의 효과가 감염율에 어떻게 영향을 받는 가를 알아보기 위한 모형설정의 과정을 구체적인 예를 통하여 살펴보았다. 자료의 성격, 구조 및 실험계획을 고려함으로써 연속적 종속모형이 설정되었고 이들 모형들은 이항 확률변수들이 독립이 아닌 많은 경우에 적용될 수 있다.

우도함수는 확률효과들의 함수들인 이항 유형의 항들을 포함하고, 모형내 모수들의 추정을 위하여 중적분을 포함한 주변우도가 유도된다. 이들 적분이 정규밀도에 관하여 행해지기 때문에, 가우스의 구적점과 관련비중을 적절히 선택한 가우스-허미트(Gauss-Hermite) 공식을 이용하여 근사적으로 적분한다. 주변 로그우도 함수로 부터의 추정방정식들이 모수들간에 비선형이기 때문에, 모형내 모수들의 최우추정치를 구하기 위하여 반복적인 연산방식이 요구된다. 심플랙스 방법이 주변 우도함수를 최대로 하기 위하여 행해졌으나, 이 방법은 2차 도함수의 Hessian 행렬을 제공하지 않는다. 최우추정치들의 근사적인 분산 공분산 행렬은 심플렉스 과정의 변형으로 계산된다. 비록 연속적 종속모형이 유행성 독감자료에 대해 전개되었지만, 이 모형은 매우 다른 실험상황에서도 유용하게 적용될 수 있다.

참고문헌

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

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

[3] 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.

[4] Coffey, M. (1988). A random effects model for inary data from dependent samples, Biometrics, Vol. 31, 737-743.

[5] Collet, D. (1991) Modelling binary data, Chapman and Hall, London.

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

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

[8] 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 Series B, Vol. 39, 1-38.

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

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

[11] Korn, E. L. and Whittemore, A. S. (1979). Methods for analyzing panel studies of acute health effects of air pollution, Biometrics, Vol. 33, 631-679.

[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.

[14] Stiratelli, R., Laird, N., and Ware, J. H. (1984). Random effects models for serial observations with binary responses, Biometrics, Vol. 40, 961-971.

A Generalized Model For Categorical Data From Epidemiological Studies[3)]

Jaesung Choi[4)]

Abstract

This paper discusses the effectiveness of an infection rate under a certain disease on an immunity rate by a protective inoculation. A sequence of dependence models concerning the infection rate is derived by defining conditionally nested binary random variables for the analysis of polytomous data with hierarchical response scale. Maximum likelihood estimates based on the marginal log-likelihood function are obtained numerically in the Nelder and Mead's(1965) simplex method.

3. This paper was supported in part by NON DIRECTED RESEARCH FUND, Korea Research Foundation, 1994.

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