Journal of Korean
Data & Information Science Society
2004, Vol. 15, No. 2, pp. 379~386

A Generalized Mixed-Effects Model for Vaccination Data

Jaesung Choi[1)]

Abstract

This paper deals with a mixed logit model for vaccination data. The effect of a newly developed vaccine for a certain chicken disease can be evaluated by a noninfection rate after injecting chicken with the disease vaccine. But there are a lot of factors that might affect the noninfecton rate. Some of these are fixed and others are random. Random factors are sometimes coming from the sampling scheme for choosing experimental units. This paper suggests a mixed model when some fixed factors need to have different experimental sizes by an experimental design and illustrates how to estimate parameters in a suggested model.

Keywords : logit, mixed model, nested design, noninfection rate, sampling unit

1. 서론

어떤 연구목적을 위해 수집된 통계적 자료의 분석방법들은 자료의 분석을 위해 모형을 고려하고 있는 가에 따라 모형에 근거한 분석방법과 모형에 근거하지 않은 분석방법들로 나눌 수 있다. 개체의 반응에 영향을 미칠 수 있는 많은 변수들이 있을 때, 반응에 영향을 미치는 변수들의 개별적인 효과의 추론 및 이들 변수들 간의 관련성을 파악하기 위한 효율적이고 체계적인 분석방법들은 일반적으로 모형에 근거하게 된다. 통계적 자료를 모형에 근거하여 분석할 때, 자료의 유형에 따라 모형들은 다양한 형태로 주어진다. 통계적 자료의 유형은 개체의 반응이 실수구간내의 모든 값을 가정할 수 있는 경우의 연속형 자료와 유한개의 범주로 구분되는 범주형 자료로 분류할 수 있다.

연속형 자료분석을 위한 선형모형 및 비선형모형에 대한 연구는 활발히 이루어져 왔으나, 범주형 자료분석을 위한 모형이용 및 개발은 상대적으로 미흡했고 최근에 들

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

어서야 컴퓨터의 급속한 발달과 함께 큰 진전을 보고 있다. 그러나 본 연구자가 관심을 갖는 임상의학 및 전염병학 분야에서 수집되는 대부분의 범주형 자료들을 분석하기 위한 모형들에 관한 연구는 여전히 미진한 상태이다.

젖소의 유선염에 대한 백신의 예방접종 및 소아의 감염성 질병에 대한 백신접종으로 인한 부작용 때문에 사회적 물의가 되고, 유행성 독감에 대한 백신의 효과에 대하여 의학자들 간에 논란이 있어 왔다. 그러나, 이러한 문제점에도 불구하고 백신접종자료를 분석하기 위한 통계적 검증방법이 마련되지 않아 이에 대한 연구가 필요하다고 본다. 감염성 질병의 예방백신에 대한 연구가 미흡한 이유는 예방의학 또는 면역학에 대한 전문적 지식이 필요하며, 백신접종자료에 관련된 많은 요인들의 복합성 때문에 자료분석에 이용될 수 있는 적절한 모형이 제시되어 있지 않기 때문이라 생각된다.

인체 또는 동물의 질병들은 감염성 여부에 따라 감염성 질병과 비감염성 질병으로 나누어진다. 본 연구는 관심질병이 한 감염성 질병이고 이 질병을 예방하기 위한 목적으로 개발된 백신의 효과를 추론하는 데 있다. 관심백신의 효과를 추론하기 위해 연구자의 조사질병에 대해 예방접종이 필요한 개체들의 모집단에서 확률표본을 추출하고, 표본으로 추출된 개체들에 대해 개발 백신으로 접종한다. 예방접종후 관심질병의 발생시기에 그 질병에 대한 감염여부를 관측하게 된다. 개체의 감염여부는 이가의 반응변수이고, 수집된 자료는 이가의 범주형 자료로 주어지게 된다. 따라서, 범주형 자료로 주어지는 백신접종자료의 분석을 위해서는 예방접종할 개체들 또는 실험단위들의 추출단계에서부터 예방접종후 자료를 수집하기 까지 많은 점들이 고려되어야 한다. 즉, 백신접종자료를 분석하기 위한 모형구축을 위해서 자료에 영향을 미칠 수 있는 외적요인들이 실험단위를 얻기위한 표본추출계획으로부터 발생하는 가, 그리고 관심백신의 효과를 추론할 때, 백신들의 제조방법, 접종방법 또는 접종용량간에 차이가 있는 가를 알아보기 위한 모형에 관한 연구는 아직 미흡한 형편이다.

개체에 대한 반응이 두 개의 값만을 취하는 이가자료로 주어지고 반응에 영향을 미치는 두 가지 유형의 요인, 즉, 고정요인(fixed factor)과 확률요인(random factor)이 존재할 때 이들이 자료에 미치는 영향을 조사하게 된다. 실험 또는 관측조사로부터 주어지는 자료들은 다양한 요인들을 포함하게 되고 이들 요인들을 모형에 포함시켜 그 효과를 추론하는 것이 때로는 간단하지가 않게 된다. 특히 반응에 영향을 미치는 확률요인이 실험계획으로부터 주어질 때 좀더 복잡한 경우의 자료분석 모형이 필요하게 된다.

범주형 자료분석과 관련된 문헌들중 Anderson and Aitkin(1985)은 조사설문지의 분석을 위한 면담자들의 변동을 다루었다. 이들은 이원지분 조사계획(two-level nested survey design)으로 발생하는 분산성분들을 추정하기 위하여 로짓모형(logit model)에 대한 최우추정법을 제시했다. Conaway(1990)는 각 개체에 대한 반복측정치들이 독립임을 의미하는 국부적 독립모형과 반응들간의 추가종속성을 반영하기 위한 종속모수들을 포함하는 종속모형들을 다루었다. Wu and Ding(1999)은 후천성 면역결핍증(Aids)환자의 임상치료에서 개별환자와 전체환자 집단에 대한 바이러스역학 모수를 추정하기 위한 계층적 비선형의 혼합효과모형을 제시하고 있다. 이와같이 여러 연구분야에서 수집되는 다양한 유형의 범주형 자료들을 분석하기 위한 방법들은 타당한 모형의 근거하에 개발되어 왔으나, 관심백신들의 예방접종과 관련한 범주형 자료를 분석하기 위한 모형의 제시는 문헌상에 제시되어 있지 않다.

본 논문에서는 실험계획과 관련된 확률요인과 처치로서의 고정요인을 고려한 혼합

모형에서의 자료분석방법을 논의하고자 한다. Im and Gianola(1988)는 이원지분계획(two-way nested design)으로부터 발생하는 분산성분들을 추정하기 위하여 이항자료에 대한 혼합모형을 다루고 있다. 개체 또는 실험단위의 반응에 대한 단순척도(pure scale)의 관측반응이 이가의 범주로 주어질 때, 실험 또는 조사로부터 수집된 자료는 이가자료를 구성하게 되고 이들이 반응범주의 도수로 표현되면 이항자료(binomial data)라 한다.

본 연구는 관심모집단의 개체에 대한 관측이 이가 범주중 하나로 관측되고 고정요인들의 효과를 추론하기 위한 실험계획에서 개체의 반응에 영향을 미치는 확률요인을 가정할 때 혼합모형의 제시와 함께 모수내 미지모수들을 추론하는 방법을 논의한다.

2. 실험환경

백신접종자료에 대한 자료분석 모형을 제시하기 위하여 다음과 같은 실험환경을 가정한다. 양계업자는 가금류의 전염병, 예를들면 뉴캣슬병, 전염성 기관지염, 가금인플루엔자등의 전염성 질병에 관심을 두게 된다. 이들 전염병의 발생은 양계농가에 치명적인 손실을 입힐 수 있기 때문이다. 한 특정전염병의 백신개발에 관심을 갖는 개발업체의 한 연구자가 그 전염병의 예방을 위해 개발중인 백신제품의 효과를 추론하는데 관심을 갖는다 하자. 개발된 백신제품에 영향을 미치는 요인들로 이용된 백신의 종류, 백신접종방법 그리고 계사의 크기등을 고려할 수 있다. 백신의 종류는 그 제조방법에 따라 사균백신, 생균백신 그리고 톡소이드등으로 구분된다. 세 종류의 백신을 A, B, C로 나타내기로 한다. 백신접종방법으로는 개체접종인 점안접종법, 집단을 대상으로 하는 음수접종법 그리고 분무접종법등을 생각할 수 있다. 이들 세가지 접종법을 I, II 그리고 III으로 나타내기로 한다. 개발중인 백신제품에 영향을 갖는 것으로 예상되는 요인들의 효과를 알아보기 위하여 이원지분계획을 가정한다. 양계농가가 전국에 흩어져 있으므로 먼저 전국의 양계지역집단에서 일부지역을 추출한 다음 추출된 지역내에서 일부 양계농가들을 추출하여 해당농가의 계군들을 대상으로 접종한다. 이때 두 가지 관심요인들의 수준들은 표본으로 추출된 지역들에 임의로 백신종류를 배정한 후, 지역내 농가들에 세가지 접종방법중 하나를 임의로 배정하는 실험을 계획한다. 이러한 실험계획은 지역간의 변동이 제품효과에 크게 영향을 미치지 않을 때, 연구자에게 단순하고 효과적인 방법일 수 있을 뿐만 아니라 체계적이고 일관성 있는 자료를 수집할 수 있으리라 판단된다.

개발중인 백신제품을 접종할 실험단위들인 닭을 추출하기 위한 이원지분계획으로부터 백신접종후 관심질병에의 비이환율에 영향을 미치는 두 가지 변동요인들이 발생하게 된다. 하나는 지역간의 차이로 인한 변동이고, 다른 하나는 지역내 농가간의 차이로 인한 변동을 생각할 수 있다. 표본으로 추출된 지역들은 지역들의 집단에서 임의로 선정되기 때문에 지역에 따른 효과는 확률효과로 간주할 수 있다. 마찬가지로 지역내 일부농가의 추출 또한 임의로 선정되므로 농가효과 또한 확률효과이다. 따라서 표본추출방법에 따른 두 요인, 즉 지역과 농가는 확률요인이다. 관심의 주된 두 요인들(백신의 종류와 백신접종방법)은 각기 세 수준을 갖는 고정요인들로 간주된다. 이러한 실험환경의 설정에 대한 주안점은 생물의 전염성질병에 대한 백신효과를 추론하기 위한 실험계획에서 백신효과에 영향을 미칠 수 있는 고정요인들중 일부는 크기가 다

른 실험단위를 취할 수 있다는 점에 착안하고 있다.

3. 모형

실험환경의 가정으로부터 모형설계를 위해 개체의 관측반응을 생각해 보자. 개체에 대한 관측반응은 개발백신의 접종후 관심질병에의 감염여부가 된다. 따라서, 감염여부는 개체의 반응변수이고 접종후 관심질병에 감염 또는 이환되었다면 0 아니면 1의 값을 갖는 이가의 반응변수 Y로 나타낸다. 개발된 또는 개발중인 백신제품을 양계농가에서 사육하고 있는 닭에 접종한 뒤 비이환율(관심질병에 감염되지 않을 확률)을 π라 둔다. 개발백신의 비이환율에 영향을 미칠 수 있는 두 고정요인의 효과를 알아보기 위해 실험단위들인 양계닭을 이원지분계획에 의해 추출하게 되므로 지역효과와 농가효과에 대한 가정이 필요하게 된다. 지역집단에서 임의로 추출된 지역 i는 비이환율에 영향을 미치는 지역효과를 갖게 되나 그 효과는 $N(0,\ \sigma_a{}^2)$를 따르는 확률효과로 간주한다. 또한 지역 i에서 임의로 추출된 양계농가 j의 효과도 $N(0,\ \sigma_f^2)$를 따르는 확률효과로 간주된다. 양계농가 j에서 추출된 닭의 수를 n_{ij}라 둔다. 지역 i, 양계농가 j에서 실험단위로 추출된 k번 째 닭에 백신접종후 조사질병에 감염되면 0, 아니면 1의 값을 갖는 이가반응변수를 Y_{ijk}라 둔다. 따라서 관심백신 제품의 비이환율에 영향을 미치는 것으로 간주되는 고정요인들과 실험단위들의 추출방법과 결부된 확률요인들이 고려된 혼합모형이 자료분석에 이용된다. $P(Y_{ijk}=1)$을 π_{ij}라 두면 자료분석을 위한 일반적인 모형은 다음과 같다.

$$g(\pi_{ij}) = \boldsymbol{x}'\boldsymbol{\beta} + \boldsymbol{z}'\boldsymbol{\alpha} \tag{3.1}$$

단, $\boldsymbol{x}' = (x_{ij1}, x_{ij2}, \ldots, x_{ijk})$로서 k개의 독립변수와 관련된 자료를 나타내는 행벡타이고, $\boldsymbol{z}' = (z_{ij1}, z_{ij2}, \ldots, zx_{ijr})$ 은 r개의 확률변수와 관련된 자료를 나타내는 행벡타이다 $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)'$는 k개의 미지모수들을 나타내는 열벡타이고, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_k)'$는 r개의 미지모수들을 갖는 열벡타이다. 여기서 $g(\cdot)$는 연결함수를 나타낸다. 실험단위를 추출하기 위한 이원지분계획에서 일부 고정요인들이 지분관계에 있는 확률요인들의 수준과 관련지어질 때, 이들 고정요인들의 효과는 관련된 확률효과들의 변동으로 그 효과를 추론하는 것이 바람직하다. 예를 들면, 대단위의 여러 계사들을 운영하는 기업형 양계농장에서는 각 계사내 닭들에 임의로 선정된 백신종류와 임의 선택된 접종방법으로 접종하는 것이 가능하지 않게 된다. 따라서, 계사간에 임의 선정된 백신종류를 이용하여 표본으로 추출된 닭에 임의 선택된 접종방법으로 접종하는 것이 다소 편리할 수 있다. 하지만, 일차 표집단위로 선택된 지역에 임의 선택한 백신종류를 이용하고 지역내 추출된 이차표집단위인 농가에 임의 선택한 접종방법을 이용함이 용이하다. 이때 고려해야 할 점은 고정요인들이 서로 다른 크기의 표집단위를 이용하게 됨으로써 야기되는 요인들의 효과를 추론하는 것이 문제

가 된다. 예방접종후 관심질병에 대한 비이환율은 정준연결함수(canonical link function)을 이용할 때, 백신접종자료에 대한 일반화된 선형모형은

$$Logit_{ijkl} = L_{ijkl} = \log[\frac{\pi_{ijkl}}{(1-\pi_{ijkl})}] = \alpha + \tau_i + a_{ij} + m_k + (\tau m)_{ik} + f_{ijkl} \qquad (3.2)$$

$i = 1, 2, \ldots, t, j = 1, 2, \ldots, a,\ k = 1, 2, \ldots, b$ 이고 $l = 1, 2, \ldots, c_j$ 이다.
단, π_{ijkl}는 지역 j 에서 백신종류 i, 지역내 농가 l 에서 접종방법 k가 행해진 후의 비이환율을 나타낸다. α는 전체평균이고, τ_i는 백신종류 i의 효과를 나타내며, m_k는 접종방법 k의 효과를 나타낸다. 그리고 $(\tau m)_{ik}$는 백신종류 i와 접종방법 k와의 교호작용을 나타낸다. 여기서 a_{ij}는 지역효과를 나타내는 확률효과이고 $N(0, \sigma_a^2)$을 따른다고 가정한다. f_{ijkl}은 백신종류 i가 이용된 지역 j내 접종방법 k가 행해진 지역내 농가 l의 확률효과를 나타내며 $N(0, \sigma_f^2)$를 따른다고 가정한다.

4. 자료 예

백신접종자료의 분석 예로 다음과 같은 연구를 가정해 본다. 양계농가에서 관심을 갖는 감염성 질병인 뉴캣슬병의 예방백신 개발에 힘쓰고 있는 백신제조업체를 생각한다. 이 업체는 개발중인 백신제품이 백신제조에 이용되는 세 가지 백신종류중 백신종류에 따른 효과가 차이가 있는 가 그리고 가능한 접종방법중 어떤 접종방법이 가장 효과적인 가를 알아보고자 한다. 전국의 양계농가지역에서 임의로 6개 농가지역을 추출한 다음 추출된 각 지역에서 임의로 3군데의 양계농가를 추출한다고 가정한다. 실험을 시행하는 계획은 세 종류의 백신 A, B 그리고 C를 동일 수의 두개 지역에 임의로 배정한다. 지역내 추출된 세 군데의 농가에 임의로 세 가지 접종방법 Ⅰ, Ⅱ 그리고 Ⅲ중 하나로 접종한다. 전염병의 발생때 예방접종후 감염되지 않은 닭의 수를 관측한다. 수집된 자료가 다음과 같다고 하자.

<표 4.1> 백신접종자료

지역	백신종류	농가	방법	반응 감염	반응 미감염
1	A	1	Ⅲ	30	220
	A	2	Ⅰ	200	300
	A	3	Ⅱ	80	320
2	B	1	Ⅰ	65	135
	B	2	Ⅲ	25	175
	B	3	Ⅱ	25	225
3	B	1	Ⅲ	15	365
	B	2	Ⅰ	50	300
	B	3	Ⅱ	20	380
4	A	1	Ⅰ	300	400
	A	2	Ⅱ	150	500
	A	3	Ⅲ	50	450
5	C	1	Ⅲ	65	315
	C	2	Ⅱ	90	210
	C	3	Ⅰ	88	132
6	C	1	Ⅱ	54	216
	C	2	Ⅰ	105	245
	C	3	Ⅲ	42	398

위 <표 4.1>의 백신접종자료에 식(3.2)를 적용해 보자. 식(3.2)의 적용은 두 고정요인 백신종류와 백신접종방법이 예방백신의 효과로 간주되는 비이환율에 영향을 미칠 수 있는 요인들일 때, 그 효과를 추론하기 위해 예방백신 접종후 뉴캣슬병에 감염되지 않은 각 지역내 농가의 비이환율을 조사한다. 이들 비이환율의 Logit 변환값들을 두 고정요인의 효과와 실험계획과 관련된 확률효과를 포함시킨 혼합효과 모형으로 자료분석하게 됨을 의미한다. 모형을 적합시킨후 모수들의 추정값들과 이들 표준오차들은 <표 4.2>와 같이 주어진다. 분산성분들의 추정값은 $\widehat{\sigma_a^2}$=0.1850 이고 $\widehat{\sigma_f^2}$=0.0077 이다. 모형에 대한 이탈도(deviance)의 값은 0.7541이고 해당하는 자유도는 7이다. 따라서 평균이탈도는 근사적으로 0.1077이 된다. 이는 인위적으로 생성된 자료가 예상보다도 상당히 변동이 적음(highly underdispersed)을 나타내고 있다. <표 4.3>은 고정효과들에 대한 Type 3 Tests의 결과를 나타내고 있다.

<표 4.2> 식(3.2)의 고정효과추정값과 표준오차

고정효과	표준오차
$\hat{\alpha}$=1.9116	0.3117
$\hat{\tau_A}$=0.8197	0.4355
$\hat{\tau_B}$=0.9316	0.4384
$\hat{m}_{\mathrm{I}}$ =-1.3243	0.0622
$\hat{m}_{\mathrm{II}}$ =-0.7555	0.0631
$\widehat{\tau m}_{A\mathrm{I}}$ =-0.3865	0.0883
$\widehat{\tau m}_{A\mathrm{II}}$ =-0.0937	0.0846
$\widehat{\tau m}_{B\mathrm{I}}$ =-0.2132	0.1050
$\widehat{\tau m}_{B\mathrm{II}}$ =0.4030	0.1254

<표 4.3> 고정효과의 Type 3 Tests

효과	분자의 자유도	분모의 자유도	F 값	Pr > F
백신종류	2	4	3.43	0.1359
방법	2	4	786.82	<.0001
백신종류*방법	2	4	9.47	0.0256

위 <표 4.3>으로부터 두 고정요인들의 교호작용은 유의수준 0.05에서 유의함을 보여주고 있다.

5. 결론

본 논문은 조류의 전염병과 관련하여 백신을 개발하고자 하는 백신업체가 개발백신에 영향을 미칠 수 있는 여러 고정요인들의 효과를 알아보고자 하는 경우에 실험단위를 추출하기 위한 표본추출계획으로 부터 표본추출단위들이 일부고정요인들의 효과를 알아보기 위한 실험단위로 이용되는 경우를 고려하고 있다. 이 경우에 서로 다른 크기의 실험단위들이 실험에 이용되기 때문에 고정요인들의 효과를 분석하기 위한 표준오차의 크기도 달라짐을 알 수 있다. 본 논문은 개발백신의 효과를 알아보기 위해 수집되는 자료가 이가의 범주형 자료이고, 개체의 반응에 영향을 미치는 요인들이 고정요인과 확률요인 둘다 포함하고 있는 경우를 가정하고 있다. 여기서 고정요인은 유한개의 수준으로 구성된 질적변수이고 확률요인은 실험단위를 얻기 위한 표본추출계획으로부터 발생하는 요인을 고려하고 있다. 예방백신 접종후 관심질병에의 비이환율에 대한 혼합요인들의 효과를 추론하기 위해 로짓을 이용한 로짓모형을 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. 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.
4. Conaway, M. R.(1990). A random effects model for binary data, biometrics, Vol. 46, 317-328.
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. Wu, H. and Ding, A. A. (1999). Population HIV-1 Dynamics in Vivo: Applicable models and Inferential Tools for Virological Data from AIDS Clinical Trials, Biometrics, Vol55, No. 2, 410-418.

[2004년 1월 접수, 2004년 5월 채택]