Journal of the Korean
Data & Information Science Society
1999, Vol. 10, No. 1, pp. 199 ~ 205

초과변동의 이항자료에 대한 혼합효과 모형[1]

최재성[2]

요약

본 논문은 초과변동을 갖는 이항자료의 효과적인 분석을 위하여 자료수집 과정에서 반응에 영향을 미치는 요인들을 고려한 혼합효과 모형을 제시하고 이탈도(deviance)를 이용하여 그 타당성을 논의하고 있다.

주제어 : 초과변동, 혼합효과, 이탈도

1. 서론

범주형 자료는 관측단위 또는 실험단위들의 하나 또는 둘 이상의 반응변수들의 관측값들이 유한개의 범주로 분류되는 자료를 의미한다. 이들 범주형 자료는 반응범주의 수에 따라 크게 두 부류로 구분할 수 있다. 하나는 두개의 반응범주를 갖는 이가자료(binary data)이고, 다른 하나는 셋 이상의 유한개수의 반응범주를 갖는 다가자료(polytomous data)로 나누어진다. 이가자료가 반응에 영향을 미치는 인자들의 수준결합에서 도수로 나타낼 때 이항자료(binomial data)라 불리어진다. 이항자료를 분석하기 위한 확률모형은 이항분포가 가정된다. 이항분포를 가정하기 위한 조건들은 첫째로 반응을 나타내는 개체 또는 관측단위간의 독립성이 보장되어야 하며 둘째로 각 개체에 대한 반응변수가 관심범주로 관측될 확률은 일정하여야 된다. 그러나 생물분석실험 또는 임상실험으로 부터 수집되는 이항자료들중 다수가 이항분포에 대한 가정을 만족시키지 않을 수 있다. 왜냐하면, 실험에 이용되는 개체들의 환경 또는 유전적 특성으로 인하여 동일 반응변수의 관심범주에 대한 확률이 같지 않을 수 있으며 또한 동물실험에서 한 어미의 새끼들로 부터 유사한 반응이 주어질 때 독립성이 의심받을 수 있다. 이항분포에 대한 가정이 만족스럽지 않을 때, 이항분포의 성질 또한 만족되지 않는다. 이 경우 자료분석을 위하여 가정된 이항분포에서 예상할 수 있는 분산보다 더 많은 변동을 초과변동(overdispersion)이라 한다. 초과변동의 이항자료를 분석하

[1]본 연구는 1998년도 계명대학교 비사연구기금으로 이루어졌음
[2](704-701) 대구광역시 달서구 신당동 1000, 계명대학교 통계학과 교수

기 위한 방법들이 많은 문헌들에서 논의되어 왔다. 이들 문헌중 독성학 실험으로 부터 주어진 자료에서 초과변동의 현상에 관한 초기의 많은 연구들이 Haseman and Kupper(1979)에 의해 조사되었다. Cox and Snell(1989)과 McCullagh and Nelder(1989)는 초과변동이 초래하게 되는 일부 원인들과 결과에 관하여 논의하고 있다. Collet(1991)는 이항분포의 초과변동을 다루기 위한 여러방법들을 비교분석하고 있다. Williams(1982a)는 초과변동을 고려한 변동모수(dispersion parameter)를 추정한후 이를 모형에 가중치로 이용하는 방법을 제시하고 있다. Follman and Lambert(1989)는 분포가 기술되지 않은 확률효과를 갖는 모형을 자료에 적합시키기 위한 비모수적 방법을 기술하고 있다. 앞서 논의된 문헌들은 하나의 확률효과를 갖는 경우에 다양한 모형제시와 함께 모형내 미지모수들을 추론하기 위한 방법들을 논의하고 있으나 본 논문은 이와는 달리 실험성격 또는 표본추출계획으로 부터 야기될 수 있는 하나 또는 두개의 확률효과들을 고려해야 하는 경우의 모형설정을 제시하고 제시된 모형내 미지모수들을 추론하는 방법을 논의하고자 한다.

2. 하나의 확률효과를 갖는 혼합효과 모형

초과변동의 이항자료를 분석하기 위해 하나의 확률효과를 포함시키는 경우를 생각해본다. 어떤 농작물의 수확에 영향을 미치는 해충을 막기위한 살충제의 독성시험에서 용량의 증가에 따른 해충의 살상율은 용량의 변화에만 의존할 때 이항자료의 초과변동은 고정효과만을 갖는 모형에서 고려된다. 그러나, 일정수의 해충들이 들어있는 집단(batch)간의 효과가 다르다고 인식될 때, 즉 실험환경의 변화나 집단간 해충의 유전적 성격이 다소 차이가 있을 때 집단간 살상율은 일정하지 않게되고 이항자료의 초과변동을 야기하는 또하나의 변동요인이 되므로 이효과를 확률효과로 모형화 하게된다. 적절한 모형의 기술을 위하여 독성시험에서의 용량의 수준들이 $i = 1, 2, ..., k$개 있다 가정하자. i번째 용량, dose_i가 살포된 해충집단i에서 살상된 해충의 수를 y_i라 두자. 이 때 y_i는 i번째 해충집단에서 dose_i가 살포되었을 때 개체의 살상률이 π_i인 해충집단에서 살상된 해충의 수 이다. i번째 해충집단에서의 해충수를 n_i라 둘 때 y_i는 모수가 n_i, π_i인 이항분포를 따른다. 이 때 수집된 자료를 분석하기 위한 모형의 형태는 다음과 같다.

$$g(\pi_i) = \beta_0 + \beta_1 \text{dose}_i + \theta_i$$

단, g는 연결함수(link function)이고 i는 k개 서로 다른 집단을 나타내며 θ_i는 $N(0, \sigma^2)$인 분포를 따른다. 이 경우 반응확률에 영향을 미치는 독립변수는 연속적인 설명변수로 간주된다. 이와는 달리 반응확률에 영향을 미치는 독립변수가 k개의 수준을 갖는 고정효과로 간주될 때 위 모형식은 다음과 같이 표현될 수 있다.

$$g(\pi_i) = \beta_0 + \text{dose}_i + \theta_i$$

모형내 미지모수들을 추론하기 위하여 $N(0,\sigma^2)$인 분포를 따르는 θ_i를 σz_i 라 두자. 그러면 위의 모형들은

$$g(\pi_i) = \beta_0 + \beta_1 \text{dose}_i + \sigma z_i$$

와

$$g(\pi_i) = \beta_0 + \text{dose}_i + \sigma z_i$$

로 표현될 수 있다. 단, z_i는 $N(0,1)$인 분포를 따르게 된다. 그러므로 모형내 미지모수들의 추론방법으로 최우법을 사용할 수 있다. 왜냐하면, i번째 용량에서 살상된 해충수 $y_i, i = 1,2,...,k$는 이항분포를 따르므로 모형내 미지모수들의 우도함수(likelihood function)는

$$L(\beta,\sigma;\mathbf{z_i}) = \prod_{\mathbf{i=1}}^{\mathbf{k}} \binom{\mathbf{n_i}}{\mathbf{y_i}} \pi_{\mathbf{i}}^{\mathbf{y_i}} (\mathbf{1}-\pi_{\mathbf{i}})^{\mathbf{n_i}-\mathbf{y_i}}$$

이다. g가 로지트 연결함수일 때

$$L(\beta,\sigma;\mathbf{z_i}) = \prod_{\mathbf{i=1}}^{\mathbf{k}} \binom{\mathbf{n_i}}{\mathbf{y_i}} \frac{(\exp\{\beta_0 + \beta_1 \text{dose}_\mathbf{i} + \sigma \mathbf{z_i}\})^{\mathbf{y_i}}}{(\mathbf{1} + \exp\{\beta_0 + \beta_1 \text{dose}_\mathbf{i} + \sigma \mathbf{z_i}\})^{\mathbf{n_i}}}$$

이다. z_i들은 $N(0,1)$인 분포를 따르는 독립인 변수들이라 가정한다. 위의 우도함수를 z_i들에 관하여 적분하여 주변우도함수를 구한후 이 주변우도함수를 최대로 하는 모수의 최우추정치를 구한다.

3. 두개의 확률효과를 갖는 혼합효과 모형

두개의 확률효과를 갖는 혼합효과 모형의 설정에 관한 논의는 문헌상에서 찾아보기가 쉽지 않다. 이러한 모형을 설명하기 위하여 다음과 같은 실험을 생각해본다. 한 제약회사에서 돼지의 특정 감염성 질병을 예방하기 위하여 면역백신을 개발하고자 한다. 우수한 면역백신의 개발을 위해 우선 항체생성률에 영향을 미치는 변수들이 무엇인가를 조사하고 이들 변수들의 효과를 알아보고자 한다. 면역백신의 항체생성률에 영향을 미치는 변수로 백신의 종류(사백신,생백신), 접종방법(경구용, 주사용) 과 접종용량의 세 가지를 생각할 수 있다고 하자. 개발중인 백신제품의 항체생성률에 영향을 미칠 수 있는 세 변수들의 효과를 파악하기 위하여 실험단위들인 돼지를 이용해야 한다. 전국에 분포한 축산농가로 부터 실험단위들인 개별가축을 추출하여 실험하는 것이 용이하지 않으므로 집락추출법을 이용하여 실험을 행하고자 한다. 육류용 돼지를 사육하고 있는 축산농가들의 지역별 집단에서 일부 지역들을 확률화의 원리에 의해 추출한 다음 추출된 지역내 축산농가들의 집단에서 일부 축산농가들을 확률표본으로 취함으로써 실험에 필요한 실험단위들인 돼지들을 얻을 수 있다. 실험단위들인 가축을 얻기 위하여 집락추출방법을 이용할 때, 지역간의 기후차이, 축산농가간의 사육방법등의 차이로 지역간, 농가간의 항체생성률에 어느정도의 변동

을 예상할 수 있다. 따라서 실험단위들의 추출방법에 따른 지역간, 농가간의 항체생성률에 대한 변이는 각 축산농가에서 관측되는 항체가 생긴 돼지의 수들에 대한 확률분포로 이항분포를 가정할 때 초과변동을 야기하는 원인이 된다. 그러므로 이러한 변동요인들을 고려한 분석모형이 마련되어야 한다. 이 에에서 생각할 수 있는 분석모형을 구체적으로 설명하기 위하여 다음과 같이 부호들을 정의한다. 백신의 종류에 따른 두 수준효과를 $\alpha_i, i = 1,2$, 백신의 접종방법에 따른 두 수준효과를 $\beta_j, j = 1,2$, 백신의 접종용량에 따른 다섯 수준효과를 $\gamma_k, k = 1,2,3,4,5$ 로 나타내며, 전체 L개 축산지역집단 $\{1,2,...,L\}$에서 l번째 지역효과를 θ_l, l번째 지역내 축산농가집단 $\{1,2,...,H_l\}$에서 h번째 축산농가의 효과를 δ_{hl}로 정의한다. 이원 지분계획(two-way nested design)하의 두 개의 확률효과를 갖는 삼원 처치구조(three-way treatment structure)를 갖는 초과변동의 이항자료를 분석하기 위한 혼합모형의 한 형태는 다음과 같다.

$$g(\pi_{ijklh}) = \mu_0 + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + \theta_l + \delta_{hl}$$

단, π_{ijklh}는 축산지역 l이 표본으로 추출되고, 축산지역 l내 축산농가 h_l가 표본으로 취해졌을 때, h_l내 돼지들에 백신 i로 접종방법 j를 이용하여 용량 k를 접종하였을 때의 항체생성률을 나타낸다. 지역효과 및 농가효과들에 대한 분포는 일반적으로 정규분포를 가정하며 서로 독립이라고 가정한다. 즉, θ_l은 $N(0,\sigma_L^2)$ 이며 δ_{hl}은 $N(0,\sigma_H^2)$ 이고, θ_l 과 δ_{hl}은 모든 (h,l)에 대해 독립이라고 가정한다. 모형내 미지모수를 추정하기 위한 방법은 두개의 확률효과를 포함하고 있기 때문에 이용가능한 상업용 프로그램은 개발되어 있지 않다. 따라서, 하나의 확률효과를 갖는 혼합모형에서의 모수추정에서와 마찬가지로 가우시안 구적점(Gaussian quadruture points)을 이용하여 근사적인 주변우도함수를 구한후 이 주변우도함수의 최우추정치를 구할 수 있는 Nelder and Mead(1965)의 심플렉스 방법을 이용한다.

4. 독성시험의 예

표 1은 어떤 화학물질의 독성시험에 따른 결과의 생성자료표이다. 임신한 쥐를 대상으로 조사물질의 서로 다른 용량이 투여되었을 때의 사산율을 조사하고자 한다. 유사환경에서 사육된 수태가 가능한 암컷 쥐들 의 집단에서 일부를 확률화의 원리에 의해 표본을 취하고 또한 비슷한 수컷들의 집단에서 표본으로 추출된 수컷을 이용하여 임신한 암컷 쥐를 대상으로 실험이 행해졌다고 하자.

자료분석에 적용될 수 있는 모형들은 다음과 같다.

$$logit(\pi_{lh}) = \beta_0 + \beta_1 x \tag{1}$$

$$logit(\pi_{lh}) = \beta_0 + \beta_1 x + \delta_{hl} \tag{2}$$

$$logit(\pi_{lh}) = \beta_0 + \beta_1 x + \theta_l + \delta_{hl} \tag{3}$$

표 1: 독성시험의 생성자료 예

수컷	암컷	용량	사산 수	총수
1	1	0.1	0	14
1	2	0.5	9	13
1	3	0.8	12	13
2	1	0.1	2	13
2	2	0.5	8	15
2	3	0.8	12	14
3	1	0.1	1	15
3	2	0.5	10	15
3	3	0.8	14	15

단, 확률효과 θ_l은 수컷간의 변동효과를 나타내고 δ_{hl}은 암컷간의 변동효과를 나타낸다. 또한 두 확률효과들은 각기 $N(0,\sigma_L^2)$ 과 $N(0,\sigma_H^2)$인 분포를 따른다고 가정한다. 모형(1)을 적용하였을 때의 이탈도(deviance)는 50.38, 모형(2)에서의 이탈도는 40.08이고, 모형(3)에서의 이탈도는 36.92로 주어진다. 각 이탈도에 해당하는 자유도는 7, 6, 그리고 5이다. 유의수준 5%하에서 자유도 1의 χ^2값은 3.84이므로 암컷간의 변동효과는 유의함을 나타내고 있으나 수컷간의 변동효과는 유의하지 않음을 보여주고 있다.

5. 결론

실험 또는 조사를 통하여 얻게되는 이항자료들은 대부분 이항분포의 가정을 만족하지 못하고 있다. 이항분포에 대한 가정이 만족되지 않을 때 자료에 나타나는 현상들은 초과변동이나 감소변동(underdispersion)이 예상되나 초과변동이 일반적이다. 초과이항변동(extra binomial variation)의 자료를 분석하기 위한 적합한 모형의 설정은 수집된 자료집단에 대한 정확한 추론을 위해 필요함이 예를 통하여 설명되고 있다. 본 논문에서는 이항자료가 수집되는 과정의 실험으로 부터 관측되는 개체들의 반응이 환경 또는 유전성으로 인해 독립성이 의심받는 경우의 모형설정을 논의하고 있다. 또한 관측단위의 표본추출이 이항자료의 초과변동에 대한 변동원인이 될 경우의 모형설정과 추론방법을 구체적으로 기술하고 있으며 모형의 적합성을 이탈도(deviance)를 이용하여 검정할 수 있음을 보여주고 있다.

참 고 문 헌

1. Collet, D. (1991). *Modelling binary data* Chapman and Hall, London.

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

3. Follman, D. A. and Lambert, D. (1989). Generalizing logistic regression by nonparametric mixing. *Journal of the American Statistical Association* **84**, 295-300.

4. Haseman, J. K. and Kupper, L. L. (1979). Analysis of dichotomous response data from certain toxicological experiments. *Biometrics* **35**, 281-293.

5. McCullagh, P. and Nelder, J. A. (1989). *Generalized llinear models* (2nd edition) Chapman and Hall, London.

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

7. Williams, D. A. (1982a). Extra-binomial variation in logistic linear models. *Applied Statistics* **31**, 144-148.

A mixed-effects model for overdispersed binomial data [3]

Jaesung Choi [4]

Abstract

This paper discusses the generalized mixed-effects model for the analysis of overdispersed binomial data. Sometimes certain types of sampling designs or genetic characters of experimental units can be regarded as factors of extra binomial variation. For such cases, this paper suggests models with one or two random effects to explain overdispersion caused by those affecting factors and shows how to test for a model adequacy based on deviance.

Key Words and Phrases: overdispersed, Mixed-effect, deciance

[3]The present research has been conducted by the Bisa Research Grant of Keimyung University in 1998.

[4]Professor, Department of Statistics, Keimyung University, 1000 Sindang-Dong, Dalseogu, Taegu 704-701, Korea.