The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.2.221
28(2), 221–230

Gamma Mixed Model to Improve Sib-Pair Linkage Analysis

Jeonghwan Kim[a] · Young Ju Suh[b] · Sungho Won[c] · Jeung Weon Nah[d] · Woojoo Lee[a,1]

[a]Department of Statistics, Inha University
[b]Department of Biomedical Sciences, College of Medicine, Inha University
[c]Department of Public Health Science, Seoul National University
[d]Department of Biostatistics, College of Medicine, Korea University

(Received March 16, 2015; Revised March 31, 2015; Accepted March 31, 2015)

Abstract

Traditionally, sib-pair linkage analysis with repeated measures has employed linear mixed models, but it suffers from the lack of power to find genetic marker loci associated with a phenotype of interest. In this paper, we use a gamma mixed model to improve sib-pair linkage analysis and compare it with a linear mixed model in terms of power and Type I error. We illustrate that the use of gamma mixed model can achieve higher power than linear mixed model with Genetic Analysis Workshop 13 data.

Keywords: Linkage analysis, gamma mixed model, identity link, hierarchical generalized linear model.

1. 서론

일반화선형모형(generalized linear model)은 반응변수가 지수족을 따르는 것을 가정한다. 정규분포, 이항분포, 포아송 분포는 각각 대칭 형태의 연속형 자료, 비율 자료, 정수 형태의 자료에 대해 지수족에 속하는 가장 널리 쓰이는 분포들이다. 그러나 이들에 비해 상대적으로 감마분포의 응용에 대해서는 덜 다루어져 왔다. 감마 분포의 제한적인 응용 사례로는 양수 값을 갖는 특정한 공학 실험자료나 중도 절단이 없는 시간자료 분석을 들 수 있는데, 특히 평균과 분산이 동시에 모형화 되는 경우 분산 쪽 모형의 모수 추정을 위해서 감마 분포에 기반한 스코어 방정식(score equation)이 사용되어 왔다 (Lee 등, 2006).

감마 분포가 상대적으로 덜 고려되고 있는 상황은 임의효과모형(random effect model)에 대해서도 마찬가지이다. Lee 등 (2006)이 소개한 다단계 일반화 선형 모형(hierarchical generalized linear model)은 변량효과가 주어졌을 때 반응변수가 지수족을 따른다고 가정하는데, 선형혼합모형(linear mixed model), 이항혼합모형, 포아송혼합모형 등에 비해 감마혼합모형은 상대적으로 덜 다루어져 왔다. 이 논문에서 우리는 감마혼합모형을 이용한 사례 연구를 진행하고자 한다. 특히, 반복측정된 형제 쌍(sib pair) 자료를 가지고 양적형질 유전자좌(quantitative trait loci; QTL)와 표지 유전자

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2013R1A1A1061332).
[1]Corresponding author: Department of Statistics, Inha University, 235 Yonghyun-Dong, Nam-Gu, Incheon 402-751, Korea. E-mail: lwj221@gmail.com

들(markers) 사이의 연관(linkage)을 연구한다. 여기서 유전자좌란 유전자가 염색체위에서 차지하는 위치를 뜻하고, 표지 유전자는 특정 유전자의 존재성을 확인해줄 수 있는 유전자를 말한다. 우리는 GAW13(Genetic Analysis Workshop 13) 에서 제공된 모의자료(problem 2) (Almasy 등, 2003)를 이용하여 감마혼합 모형이 기존에 사용되어오던 혼합선형모형에 비해 검정력에 있어서 강점을 가질 수 있음을 사례연구를 통해 살펴볼 것이다.

전통적으로 연관 분석에 있어서 가장 기본이 되는 Haseman-Elston(HE) 모형은 Haseman과 Elston (1972)에서 제안되었고, 이는 형제 쌍을 이용한 연관 분석(linkage analysis)을 통해 많은 질병의 유전적 배경을 이해할 수 있게 해주었다. 먼저 i번째 형제 쌍의 관측된 표현형의 값을 각각 P_{i1}과 P_{i2}라 하자. 많은 연구에서 P_{i1}과 P_{i2}는 이변량 정규분포를 따른다고 가정되고 있다 (Almasy와 Blangero, 1998; Kruglyak과 Lander, 1995). 여기서 P_{i1}과 P_{i2}는 편의상 평균이 미리 보정된 것으로 간주하여 기댓값이 0이라고 가정하겠다. Haseman과 Elston (1972)에서 사용되었던 반응변수는

$$Y_i = (P_{i1} - P_{i2})^2$$

이고, 이 값을 연구 대상이 되는 유전자(gene locus)에서의 identical-by-descent(IBD) 수의 비율 값인 π_i에 단순회귀분석 하여 연관 분석을 시행하였다. 여기서 IBD 수란 형제쌍이 동일한 조상으로부터 물려받아 함께 가지고 있는 대립유전자(allele) 수를 뜻한다. 이때, HE 모형은

$$E(Y_i) = \beta_0 + \beta_1 \pi_i \tag{1.1}$$

으로 주어진다. 먼저 $\mathrm{Var}(P_{i1}) = \mathrm{Var}(P_{i2}) = \phi$라고 가정할 때,

$$E(y_i) = E(P_{i1} - P_{i2})^2 = 2\phi(1 - \mathrm{Cor}(P_{i1}, P_{i2}))$$

임을 주목한다. 그리고, 모집단에서 한 쌍의 사람들의 표현형이 갖는 상관계수의 평균을 r이라 하고, Q를 표현형의 분산이 해당 유전자좌에 의해 설명되는 비율, 그리고 모집단에서의 π_i의 평균값을 π_0라 할 때, 유전학적으로 다음의

$$\mathrm{Cor}(P_{i1}, P_{i2}) = r + Q(\pi_i - \pi_0).$$

이 선형 관계식을 도출할 수 있다. 이 관계식을 이용하면, 결국 (1.1)를 통해서 β_1이 연구 대상인 유전자좌에 설명되는 분산의 크기의 두 배에 반대부호에 대응된다는 사실을 알 수 있다 (Kruglyak과 Lander, 1995; Barber 등, 2004). 따라서 이때 귀무가설은 해당 유전자좌가 유전적으로 연관성을 가지고 있지 않다가 되고 이는 $H_0 : \beta_1 = 0$을 통해 검정된다. 대립가설은 $H_0 : \beta_1 < 0$이 되므로, 단방향 가설검정의 형태를 통해 연관분석을 시행하게 된다.

그 동안 HE 모형에 대해 가장 관심 깊게 다루어진 이슈 중 하나는 검정력에 대한 부분이다. 그와 관련하여 Barber 등 (2004)는 Haseman과 Elston (1972)이 사용한 정규분포에 기반한 최소제곱법(ordinary least squares)이 잘못된 분포가정에 기반하고 있기 때문에 불필요하게 분산이 커지게 되고 따라서 검정력이 떨어질 수 있음을 강조하였다. 대신에 Barber 등 (2004)는 $Y_i = (P_{i1} - P_{i2})^2$에 감마 분포를 가정하는 것이 더 타당하다고 하였고, 감마 분포를 사용하면 최소제곱법을 사용할 때에 비해 β_1의 추정치의 분산이 작아질 수 있음을 간단한 시뮬레이션을 통해 살펴보았다.

한편, HE 모형은 다양한 방향으로 확장이 시도되었는데, 가장 주목할 만한 것은 먼저 Suh 등 (2003)은 같은 부모로 부터 나온 형제 쌍 사이의 상관성을 설명하는 변량효과를 포함하는 혼합 모형을 HE 모형의 확장으로 사용하였고, Won 등 (2006)은 경시적 가계(pedigree) 자료의 연관분석을 위한 일반적인 선형 혼합 모형을 제안하였다. 그러나 반복측정된 자료에 감마분포를 연결하여 살펴보는 작업은 그 동안 이

루어지지 않았다. 본 연구에서는 감마혼합 모형을 이용하여 반복측정된 형제 쌍 연관 분석 자료를 어떻게 분석할 수 있는지를 사례연구를 통해 살펴보고자 한다.

2. 반복측정이 없을 때 연관분석을 위한 모형

먼저 반복측정이 없는 경우 연관분석에 사용되는 통계 모형에 대해서 살펴보도록 하겠다.

Kruglyak과 Lander (1995)에서는 Y_i의 분포로 무조건적으로 정규분포가 가정되어 분석되는 것에 문제가 있음을 지적하였다. 이와 같은 선상에서, Barber 등 (2004)은 $E(Y_i) = \mu_i$라 할 때, P_{i1}과 P_{i2}에 이변량 정규분포를 따르면

$$\mathrm{Var}(Y_i) = \mathrm{Var}\left((P_{i1} - P_{i2})^2\right) \propto \mu_i^2$$

임을 주목하였다. 따라서 Y_i에 대해 평균과 분산이 독립적인 함수관계를 갖는 정규분포를 가정하는 기존의 방법과는 달리, 분산이 평균의 제곱에 비례하는 함수관계를 갖는 감마분포의 사용을 지지하였다. 그런데 평균에 대한 회귀분석 모형은 여전히 (1.1)의 형태를 가지게 되므로, 항등 연결함수(identity link function)을 갖는 감마모형을 사용해야 함을 의미한다. 평균에 대한 모형은 같고, 평균-분산 관계가 다른 경우이므로, Y_i에 정규분포가 가정되든 감마분포가 가정되든 모수의 추정의 일치성(consistency)은 양쪽 모두 가지고 있으나 평균-분산 함수 관계를 올바르게 사용하게 되면 더 작은 표준오차(standard error)를 얻을 수 있고 따라서 통계량이 더 높은 검정력을 가지게 된다. Barber 등 (2004)는 Y_i의 이분산성이 더 강하게 나타날수록 감마분포를 가정한 모형에서 최소제곱 추정량(ordinary least square estimator)에 비해 검정력의 이득이 크게 나타남을 간단한 시뮬레이션 연구를 통해 확인하였다.

항등 함수를 갖는 감마 일반화 선형 모형을 사용하는 경우 통계계산에서의 유의점에서 살펴볼 필요가 있다. 일반화 선형 모형의 모수 추정 알고리즘인 반복재가중최소제곱법(iterative reweighted least squares)을 감마 일반화 선형모형이 정준 연결함수나 로그 함수일 때 에 대해서는 알고리즘이 안정적인 반면, 항등 연결함수에 대해서 그러하지 못함은 관측된 피셔 정보량 행렬을 통해 확인해 볼 수 있다. 먼저 n개의 y_i가 독립적으로 평균 μ_i와 산포 모수 ϕ를 갖는 감마 분포에서 생성되었다고 하자. 그러면 평균 μ_i와 산포모수 ϕ의 관점에서 모수화된 감마분포의 로그 가능도 함수 ℓ을 살펴보면

$$\ell(\mu, \phi; y_1, \ldots, y_n) = \sum_{i=1}^{n}\left[\left\{y_i\left(-\frac{1}{\mu_i}\right) - \log(\mu_i)\right\}\Big/\phi - \frac{\log(\phi)}{\phi} + \left(\frac{1}{\phi} - 1\right)\log(y_i) - \log\left(\Gamma\left(\frac{1}{\phi}\right)\right)\right]$$

이 된다. 먼저 정준연결함수의 경우 $1/\mu_i = x_i^T\beta = \sum_{j=1}^{p} x_{ij}\beta_j$임을 주목하자. 여기서 x_i는 설명변수 벡터이고, β는 대응되는 회귀계수이다. 이때 관측된 피셔 정보량 행렬은

$$-\frac{\partial^2 \ell}{\partial\beta\partial\beta^T} = X^T W^C X$$

이 된다. 여기서 $X = (x_1, x_2, \ldots, x_n)^T$이고 W^C는 대각행렬인데, i번째 성분이 $W_{ii}^C = 1/(x_i^T\beta)^2$이다. 모든 i에 대해 $x_i^T\beta \neq 0$라고 한다면, W^C의 모든 대각성분이 양수가 되어 위의 관측된 피셔 정보량 행렬은 양정치(positive definite) 행렬이 된다. 따라서 로그 가능도 함수는 β에 대해 오목함수(concave function)가 되고, β의 최대가능도추정량을 쉽게 찾을 수 있다.

이제 로그 연결함수의 관측된 피셔 정보량 행렬에 대해 살펴보면

$$-\frac{\partial^2 \ell}{\partial\beta\partial\beta^T} = X^T W^L X$$

Table 2.1. The number of convergence of gamma GLM with identity link

Cases	Using default starting values	Using normal models
Success	64	100
Fail	36	0

이고 W^L는 i번째 성분이 $y_i \exp(-x_i^T\beta)$인 대각행렬이다. 모든 $y_i > 0$이면, W^L의 모든 성분은 양수가 되어 관측된 피셔 정보량 행렬은 양정치 행렬이 된다.

그러나 항등연결함수를 사용하게 되는 경우에는

$$-\frac{\partial^2 \ell}{\partial\beta\partial\beta^T} = X^T W^I X$$

이고, W^I는 i번째 성분이 $2y_i/(x_i^T\beta)^3 - 1/(x_i^T\beta)^2$이 된다. 이 경우는 위의 경우와 다르게, 작은 y_i가 관측된다면 W_{ii}^I는 음수가 되어, 관측된 피셔 정보량 행렬이 양정치 행렬이 아닐 수 있게 된다. 따라서, 추정치를 구할 때 계산의 불안정성이 발생될 수 있으므로, 항등함수를 갖는 감마모형을 사용하는 경우 초기치의 선택이 수렴결과에 매우 중요한 영향을 줄 수 있다. 이것을 확인하기 위해, 우리는 x_i는 0과 1사이에서 균일분포에서 얻어지고, y_i는 평균모형이 $\mu_i = 2 - 0.5x_i$인 감마분포로부터 100개의 관측치를 만들어 R의 glm 함수를 이용하여 적합하여 보았다. 디폴트로 지정된 방법과 y_i에 정규분포를 가정하여 적합시킨 결과를 초기치로 사용한 경우를 Table 2.1에 비교하였다. 100번의 시뮬레이션을 통해 몇 번의 수렴된 결과를 얻었는지를 보고하였다.

항등 연결함수를 갖는 감마모형은 초기치에 따라 수렴의 성공률이 64%에서 100% 바뀌는 것을 확인할 수 있었다. 즉, 간단한 회귀모형이라도 감마 모형에서 항등함수를 사용하게 되는 경우 통계계산의 관점에서 알고리즘의 안정성에 대한 이슈가 발생할 수 있다는 점을 강조할만하다. 이러한 논의는 뒤에서 살펴볼 항등 함수를 사용하는 감마혼합모형의 계산에 있어서 발생하는 어려움에 유사하게 적용될 수 있다.

3. 반복 측정이 있을 때 연관분석을 위한 모형

각 형제 쌍의 표현형의 값이 반복되어 관측되는 경우 모형을 통해 관측값 사이의 상관성을 설명하는 것이 바람직하다. 이러한 상관성을 설명하는 다양한 방식이 있으나, 본 연구에서는 연관 분석에서 널리 사용되고 있는 혼합모형을 통해 살펴볼 것이다. 특히, 2절에서 소개된 정규분포를 가정한 선형 모형과 감마 일반화 선형 모형의 확장된 형태로 선형 혼합 모형과 감마 혼합모형을 순서대로 살펴보도록 하겠다.

3.1. 선형 혼합 모형

반복측정된 자료에 대한 HE 모형의 반응변수는, 반복 횟수를 말해주는 지표 j를 아랫첨자에 추가하여,

$$Y_{ij} = (P_{ij1} - P_{ij2})^2$$

로 정의하는 것에서 시작할 수 있다. 여기서 P_{ijk}은 i번째 형제쌍의 k번째 형제에게 j번째 반복에서 얻어진 값을 의미한다. $j = 1$만 가능할 때 Y_{i1}를 Y_i라 할 수 있으므로 2절에서 소개된 반복이 없는 경우의 반응변수가 된다.

이러한 모형을 분석하기 위해 선형 혼합 모형을 사용할 수 있다. 본 연구에서 사용되는 선형 혼합 모형은 i번째 형제 쌍의 변량효과 v_i가 주어졌을 때 Y_{ij}는 독립이고 $E(y_{ij}|v_i) = \mu_{ij}$이고 $\mathrm{Var}(y_{ij}|v_i) = \sigma^2$으

로 조건부 평균과 분산사이에는 함수관계가 없음을 가정하고 있는 것에 주목해야한다. 정규분포 가정하에서 다음과 같이

$$y_{ij} = x_{ij}^T\beta + v_i + \epsilon_{ij}$$

으로 표시하는 것이 가능하고, 여기서 x_{ij}는 관심대상인 유전자좌에서의 IBD 값, β는 고정효과, ϵ_{ij}는 독립이고 평균 0, 분산이 ϕ인 정규분포를 따르는 것으로 본다.

이 모형은 i번째 형제 쌍에 대해서 n_i번 반복관측이 있다고 가정되었을 때,

$$y_i = \begin{pmatrix} y_{i1} \\ \vdots \\ \vdots \\ y_{in_i} \end{pmatrix} = \begin{pmatrix} x_{i1}^T \\ \vdots \\ \vdots \\ x_{in_i}^T \end{pmatrix} \beta + \begin{pmatrix} v_i \\ \vdots \\ \vdots \\ v_i \end{pmatrix} + \begin{pmatrix} e_{i1} \\ \vdots \\ \vdots \\ e_{in_i} \end{pmatrix}$$

이 되어, 형제 쌍 내의 반복측정된 자료에 대해서

$$\mathrm{Cov}(y_i) = \begin{pmatrix} \rho & \rho & \cdots & \cdots & \rho \\ \rho & \cdots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \cdots & \rho \\ \rho & \cdots & \cdots & \rho & \rho \end{pmatrix} + \begin{pmatrix} \phi & 0 & \cdots & \cdots & 0 \\ 0 & \phi & 0 & \vdots & \vdots \\ \vdots & 0 & \ddots & 0 & 0 \\ \vdots & \vdots & 0 & \phi & 0 \\ 0 & \cdots & 0 & 0 & \phi \end{pmatrix}$$

이 되어 compound symmetric한 공분산을 가정한 모형이 된다. 여기서 $\rho = \mathrm{Var}(v_i)$이다. 이와 같은 모형의 추론은 주변가능도(marginal likelihood)가 명시적 형태(explicit form)를 기반으로 하며, R 또는 SAS와 같은 일반적인 소프트웨어를 통해 계산이 가능하다.

3.2. 감마혼합모형

Barber 등 (2004)의 논증을 반복 측정된 자료로 확장해보자. 변량효과 v_i가 주어졌을 때, j번째 시간에서 얻어진 자료 P_{ij1}과 P_{ij2}는 이변량 정규분포를 따르고, 서로 다른 시점에서 얻어진 값들과는 조건부 독립이라고 가정하자. 그러면, j번째 반복측정된 자료의 조건부 평균을 $E(Y_{ij}|v_i) = \mu_{ij} = x_{ij}^T\beta + v_i$라 할 때 조건부 분산은

$$\mathrm{Var}\,(y_{ij}|v_i) \propto \mu_{ij}^2$$

임을 알 수 있다. 따라서 Barber 등 (2004)의 연장선상에서 보면 반복측정된 형제 쌍 연관분석에 감마 혼합 모형이 사용되는 것이 정당화될 수 있다.

이와 같은 모형을 적합하기 위해서 우리는 Lee 등 (2006)의 다단계 일반화 선형 모형을 사용할 수 있다. 이번 절에서 먼저 다단계 일반화 선형 모형에 대하여 간단히 검토하고, 그것의 특별한 경우로 연관분석을 위한 감마 혼합 모형에 대해서 살펴보도록 하겠다.

다단계 일반화 선형 모형은 두개의 널리 사용되고 있는 통계 모형인 일반화 선형 모형(generalized linear model)과 혼합 선형 모형을 통합한 모형으로 다음의 두 주요 성분으로 정의된다.

- 변량효과 v가 주어져 있을 때 반응변수 y는 지수족을 따르며,

$$E(y|v) = \mu$$

일때 연결함수 g에 대해

$$g(\mu) = X\beta + Zv$$

형태로 주어진다. X는 고정효과 β 쪽의 모형 행렬, Z는 변량 효과 v 쪽의 모형 행렬이다. 주목할 점은 변량효과가 고정효과에 가법(additive) 형태로 들어간다는 점이다.

- 변량효과의 분포는 임의의 지수족 분포를 따르는 것으로 본다.

이 모형은 Breslow와 Clayton (1993)의 일반화 선형 혼합 모형(generalized linear mixed model)을 v에 정규분포가 가정된 특별한 경우로 포함한다. 고정효과 β와 변량효과 v의 추론을 위해서는 다단계 가능도(hierarchical likelihood) 함수를 사용하는데, 이는 관측된 자료 y와 변량효과 v의 결합 분포에 기반한 가능도 함수로

$$h = \log f_{\beta,\phi,\lambda}(y, v) = \log f_{\beta,\phi}(y|v) + \log f_{\lambda}(v)$$

의 관계를 갖는다. 여기서 $f(\cdot)$는 일반적으로 해당 변수의 확률 밀도 함수(또는 확률 질량 함수)를 나타내며, λ는 변량효과의 분산 성분(variance component)에 해당한다. v의 추정량은 h를 최대화하여 얻어지는데, 이것은 $\log f(v|y))$를 최대화하는 v와 같으므로 가장 확률값을 크게 갖는 변량효과 추정치를 사용하는 것이 된다. 우리는 이것을 $\widehat{v}$으로 나타낸다. β와 산포모수인 ϕ, λ를 추정하는 것은 각각

$$P_v(h) = \left[h - \frac{1}{2}\log\left| -\frac{\partial^2 h}{\partial v \partial v^T} \right| \right]_{v=\widehat{v}}$$

과

$$P_{v,\beta}(h) = \left[h - \frac{1}{2}\log\left| -\frac{\partial^2 h}{\partial (v,\beta) \partial (v,\beta)^T} \right| \right]_{v=\widehat{v},\beta=\widehat{\beta}}$$

를 최대화는 값을 사용한다. $P_v(h)$는 주변부 가능도 $\int \exp(h) dv$에 대한 라플라스 근사(Laplace approximation)에 해당하며 $P_{v,\beta}(h)$는 위의 라플라스 근사에 고정효과 β의 추론에 의해서 생기는 가능도 추정량의 편의를 보정하는 방법으로 제안된 수정된 가능도이다. 경우에 따라 $P_{v,\beta}(h)$의 계산이 복잡한 경우에는 산포모수의 추정을 위해서 $P_v(h)$의 사용도 가능하며, 이때에는 분산 성분을 주변부 가능도에서 추론하는 것이 된다. 특히, 선형 혼합 모형에 대해서 $P_v(h)$는 β에 대해 정확히 주변부 가능도 함수에서 얻어진 최대 가능도 추정량을 주며, $P_{v,\beta}(h)$는 산포 모수와 분산 성분에 대해 restricted maximum likelihood estimator(REML)를 제공한다.

v, β와 산포모수인 σ^2과 λ의 추정량을 얻어내는 목적 함수가 달라 일견 계산이 복잡해 보이지만, 이는 일반화 선형 모형에서 사용되고 있는 반복재가중최소제곱법 여러 개를 서로 결합하여 전체 추정량을 얻을 수 있음이 알려져 있다. 이러한 계산 구조는 다단계 일반화 선형 모형의 더욱 다양한 확장을 가능하게 하는데, 예를 들어, 산포모수를 다른 설명변수로 모형화 하는 것, 변량효과의 분산 성분에 다시 변량효과를 도입하여 모형이 이상치에 강건(robust)하게 만드는 것 등이 가능하게 된다. 위의 목적함수를 최대화하여 얻어진 추정량이 Breslow와 Clayton (1993)의 최대 벌점 준 가능도 추정량(maximum penalized quasi likelihood estimator)보다 일반적으로 우수하고, 특히 이항자료의 경우 훨씬 편의가 적은 것으로 알려져 있다. 자세한 내용은 Lee 등 (2006)을 참고하면 된다.

연관분석을 위해 사용되는 감마 혼합 모형의 가장 큰 특성은 앞서 보았듯이 감마 분포에 흔히 사용되는 로그 연결함수나 정준 연결함수인 역 연결함수를 사용하지 않고, 항등함수를 사용하는 것에 있다. 따라서 본 연구에서 사용하게 되는 모형은 v_i가 주어졌을 때 y_{ij}가 독립이 되고, 평균이

$$\mu_{ij} = E(Y_{ij}|v_i) = x_{ij}^T\beta + v_i$$

으로 주어진 것이 된다.

4. 감마 혼합 모형을 통한 사례 분석

본 연구에서 사용되는 자료는 GAW13 모의자료에서 주어진 세 개의 데이터셋을 형제 쌍 분석에 적합하도록 가공한 것이다. 각 데이터 셋에서 분석대상인 반응변수는 각 형제 쌍의 수축기 혈압(systolic blood pressure)으로 5개의 다른 시점에 반복측정되어 얻어졌다. 부모세대인 코호트 1과 자식세대인 코호트 2에서의 관측시간이 각각 다르므로, 코호트 2에 해당하는 자녀를 기준으로 하여 5개의 시간으로 자료를 정리하였다. 혼합 모형에 적용하기에 앞서, 이 반응변수는 각 시점에서 수축기 혈압에 영향을 주는 유의한 인자들인 총 콜레스테롤(total cholesterol), 하루에 피우는 흡연량(cigarettes per day), 공복혈당치(fasting glucose), 고혈압(high blood pressure), 고혈압 치료(hypertensive treatment) 여부, 체중 그리고 성별에 대해 보정되었다. 즉, 수축기 혈압에 위의 변수들로 회귀분석 된 다음 얻어진 잔차가 앞서 소개된 혼합모형에서의 반응변수 P_{ijk}로 사용된다. 그리고, GAW13에서 알려진 수축기 혈압과 관련이 있는 6개의 유전자좌에 대한 IBD 자료와, 이와 관련이 없는 6개 유전자좌에 대한 IBD 자료가 주어져 있다. 이 12개의 유전자좌의 이름은 아래와 같다.

- b34, b35, b36, s10, s11, s12: GAW13에서 제공된 수축기 혈압과 연관된 유전자좌
- b5, b14, b16, b18, b21, b23: 위의 6개의 유전자좌와 서로 다른 염색체(chromosome) 상에 위치한, 랜덤하게 선택된 6개의 유전자좌

각 데이터 셋은 330개의 가계(pedigree)로 구성되어 있으며, 형제 쌍 분석을 위해서 각 가계로 부터 형제 쌍을 임의로 하나씩 뽑아 사용하기로 하였다. Barber 등 (2004)에서 지적된 것처럼, 연관분석에서 통계량의 평가를 위해서는 검정력과 제 1종 오류를 고려해야 한다. 따라서 선형 혼합 모형과 감마 혼합 모형를 비교함으로써, 우리는 b34, b35, b36, s10, s11, s12에 대해서는 검정력이 높은지를 확인하고, b5, b14, b16, b18, b21, b23에 대해서는 제 1종 오류가 낮게 유지되고 있는지를 확인할 것이다.

세 데이터셋에 대한 자료분석 결과는 Table 4.1–4.3에 요약되어 있다. 선형혼합모형은 R 패키지 lme4에서 라플라스 근사를 사용하고 있는 glmer로 적합되었고 REML을 사용하였으며, 감마 혼합모형은 같은 함수로 적합되었으며, 항등연결함수인 경우 현재 REML이 지원되지 않으므로, ML 옵션이 사용되었다. 표의 결과는 12개의 고정효과에 대한 단측검정이 다중검정 형태를 가지고 있으므로, 본페로니 수정법을 이용하여 유의성을 판단하였다. 즉, 각 테이블에서 t-value ≤ -2.64로 얻어진 것만을 유의한 결과로 판단하였으며, 유의한 경우 해당 표의 유의확률 컬럼에서 "*"로 표시하였다. 분석결과를 요약하면 다음과 같다.

- 첫 번째 표에서 관련있는 6개의 유전자좌와 관련하여 선형 혼합 모형은 하나도 유의한 결과를 주지 않은 반면, 감마 혼합 모형은 4개의 유의한 결과를 주었다. 두 번째와 세 번째 표에서도 유의하게 나타난 유전자좌의 종류는 다르지만 유사한 결과가 나왔다. 즉, 위의 세 데이터셋에 대해서는 감마 혼합 모형이 훨씬 검정력이 높은 것으로 나타나고 있다.

Table 4.1. Parameter estimates for fixed effects (Bonferroni p-value cutoff for the one-sided test = 0.0042)

Variable	Normal mixed model				Gamma mixed model			
	Estimate	Std.error	t-value	p-value	Estimate	Std.error	t-value	p-value
(Intercept)	147.6240	20.3261	7.2628	0.0000	153.1629	16.1854	9.4630	0.0000
b34	−12.0664	5.8786	−2.0526	0.0401	−12.8845	3.7254	−3.4586	0.0005*
b35	−12.6191	5.9818	−2.1096	0.0349	−12.9748	3.7295	−3.4790	0.0005*
b36	−12.8725	5.8463	−2.2018	0.0277	−11.9701	3.4181	−3.5020	0.0005*
s10	−12.8965	6.0833	−2.1200	0.0340	−14.7650	3.6410	−4.0552	0.0001*
s11	2.7527	5.6112	0.4906	0.6237	0.7729	3.4291	0.2254	0.8217
s12	−0.6160	6.1227	−0.1006	0.9199	−1.6929	3.8916	−0.4350	0.6635
b5	1.5940	5.7598	0.2767	0.7820	1.0742	3.6427	0.2949	0.7681
b14	4.3868	6.1946	0.7082	0.4788	2.7060	3.8272	0.7070	0.4795
b16	4.8637	6.3461	0.7664	0.4434	3.3648	3.4160	0.9850	0.3246
b18	0.9321	6.4592	0.1443	0.8853	1.8617	4.4710	0.4164	0.6771
b21	−4.7397	6.3196	−0.7500	0.4533	−0.9110	4.0302	−0.2260	0.8212
b23	4.9718	6.1378	0.8100	0.4179	3.6966	3.7802	0.9779	0.3281

Table 4.2. Parameter estimates for fixed effects (Bonferroni p-value cutoff for the one-sided test = 0.0042)

Variable	Normal mixed model				Gamma mixed model			
	Estimate	Std.error	t-value	p-value	Estimate	Std.error	t-value	p-value
(Intercept)	173.8070	19.5839	8.8750	0.0000	174.1648	5.0283	34.6369	0.0000
b34	−13.8891	5.4001	−2.5720	0.0101	−14.7395	2.9423	−5.0095	0.0000*
b35	−12.4207	5.8215	−2.1336	0.0329	−9.2152	2.9862	−3.0859	0.0020*
b36	−13.6534	5.3207	−2.5661	0.0103	−12.4916	2.5429	−4.9124	0.0000*
s10	−10.0949	5.7791	−1.7468	0.0807	−5.5250	2.8021	−1.9717	0.0486
s11	−6.9164	5.3172	−1.3008	0.1933	−8.8242	2.5799	−3.4203	0.0006*
s12	−5.8547	6.0430	−0.9689	0.3326	−7.2768	3.0197	−2.4097	0.0160
b5	−1.4504	5.3140	−0.2729	0.7849	0.0793	2.6574	0.0298	0.9762
b14	0.6034	6.0050	0.1005	0.9200	−4.5698	2.8591	−1.5983	0.1100
b16	−3.6613	6.0934	−0.6009	0.5479	−5.4364	3.0335	−1.7921	0.0731
b18	−7.7964	6.2247	−1.2525	0.2104	−4.7434	3.5244	−1.3459	0.1783
b21	−0.7374	5.7772	−0.1276	0.8984	1.9905	3.1865	0.6247	0.5322
b23	4.6354	5.8830	0.7879	0.4307	−0.7922	2.7440	−0.2887	0.7728

- 관련없는 6개의 유전자좌와 관련하여서는 유의하게 나타나는 것이 하나도 없었다. 따라서 제 1종 오류는 양쪽 모두 잘 조절하고 있는 것으로 보인다.

따라서 본 결과는 반복측정된 형제 쌍 연관 분석에 있어서 감마 혼합 모형의 사용을 통해 상당한 검정력의 이득을 얻을 수 있음을 보여준다.

5. 결론

본 연구에서는 그 동안 상대적으로 덜 관심을 받아온 감마 혼합 모형의 한 사례연구를 진행하였다. 특히 연관 분석과 관련하여 항등 연결함수가 쓰이는 경우에 대해서 자세히 살펴보았다. GAW13 자료를 이용하여, 반복측정된 형제 쌍 연관분석에 있어서 감마 혼합 모형이 검정력을 높여줄 수 있음을 확인할 수 있었다.

Table 4.3. Parameter estimates for fixed effects (Bonferroni p-value cutoff for the one-sided test = 0.0042)

Variable	Normal mixed model				Gamma mixed model			
	Estimate	Std.error	t-value	p-value	Estimate	Std.error	t-value	p-value
(Intercept)	141.7080	21.4720	6.5997	0.0000	145.0641	12.3557	11.7406	0.0000
b34	−14.1839	5.9598	−2.3799	0.0173	−15.5772	3.3995	−4.5822	0.0000*
b35	−9.5619	6.0561	−1.5789	0.1144	−9.9911	3.3987	−2.9397	0.0033*
b36	−14.9817	5.9470	−2.5192	0.0118	−13.7309	3.4119	−4.0245	0.0001*
s10	−11.2011	6.0447	−1.8531	0.0639	−12.7578	3.2992	−3.8669	0.0001*
s11	−5.8765	5.7551	−1.0211	0.3072	−6.2093	3.1733	−1.9567	0.0504
s12	−1.8410	6.3998	−0.2877	0.7736	−1.0620	4.0484	−0.2623	0.7931
b5	1.1334	5.7406	0.1974	0.8435	0.3490	3.2504	0.1074	0.9145
b14	5.4078	6.1058	0.8857	0.3758	6.6143	3.4274	1.9298	0.0536
b16	5.4912	6.5277	0.8412	0.4002	6.4286	4.7089	1.3652	0.1722
b18	10.5346	6.1706	1.7072	0.0878	12.1011	3.9576	3.0577	0.0022
b21	3.8539	5.9819	0.6443	0.5194	3.2779	3.3156	0.9886	0.3228
b23	4.2727	6.0679	0.7041	0.4813	0.4428	3.6707	0.1206	0.9040

항등 연결함수를 갖는 감마 혼합 모형의 사용의 이득을 정량적으로 평가하기 위해서는 시뮬레이션 연구가 필요하다. 시뮬레이션 연구에서는 보통 수백 번의 반복이 필요한데, 이를 위해서는 알고리즘의 안정성이 필수적이다. 그러나 앞서 설명되었듯이 항등 연결함수를 갖는 감마 모형의 경우 피셔 정보량의 양정치성이 종종 깨질 수 있기 때문에 알고리즘의 안정성이 문제가 되므로, 이를 개선하는 알고리즘에 대한 연구가 앞으로 필요하다고 생각한다.

References

Almasy, L., Amos, C. I., Bailey-Wilson, J. E., Cantor, R. M., Jaquish, C. E., Martinez, M., Neuman, R. J., Olson, J. M., Palmer, L. J., Rich, S. S., Spence, M. A. and MacCluer, J. W. (2003). Genetic analysis workshop 13: Analysis of longitudinal family data for complex diseases and related risk factors, *BMC Genetics*, **4**(Supple 1), S1.

Almasy, L. and Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees, *American Journal of Human Genetics*, **62**, 1198–1211.

Barber, M. J., Cordell, H. J., MacGregor, A. J. and Andrew, T. (2004). Gamma regression improves Haseman-Elston and Variance Components Linkage Analysis for Sib-Pairs, *Genetic Epidemiology*, **26**, 97–107.

Breslow, N. E. and Clayton, D. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association*, **88**, 9–25.

Haseman, J. K. and Elston, R. C. (9172). The investigation of linkage between a quantitative trait and a marker locus, *Behavior Genetics*, **2**, 3–19.

Kruglyak, L. and Lander, E. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits, *American Journal of Human Genetics*, **57**, 439–454.

Lee, Y., Nelder, J. A. and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via h-Likelihood*, Chapman and Hall/CRC.

Suh, Y., Park, T. and Cheong, S. (2003). Linkage analysis of longitudinal data, *BMC Genetics*, **4**(Suppl I):S27.

Won, S., Elston, R. C. and Park, T. (2006). Extension of the Haseman-Elston regression model to longitudinal data, *Human Heredity*, **61**, 111–119.

감마 혼합 모형을 통한 반복 측정된 형제 쌍 연관 분석 사례연구

김정환[a] · 서영주[b] · 원성호[c] · 나정원[d] · 이우주[a,1]

[a]인하대학교 통계학과, [b]인하대학교 의과대학, [c]서울대학교 보건대학원, [d]고려대학교 의학통계학과

(2015년 3월 16일 접수, 2015년 3월 31일 수정, 2015년 3월 31일 채택)

요 약

전통적으로 반복 측정된 형제 쌍 연관 분석에서는 선형 혼합 모형이 사용되어 왔다. 그러나 그 모형은 관심있는 표현형과 연관된 유전자좌를 찾는 것에 있어서 검정력이 문제가 되는 것으로 지적되어 왔다. 본 연구에서 우리는 이러한 검정력 문제를 개선하는 방법으로 감마 혼합 모형을 고려하였고, 검정력과 제 1종 오류의 관점에서 선형 혼합 모형과 성능을 서로 비교하여 보았다. Genetic Analysis Workshop 13에서 제공된 자료를 이용하여 살펴본 결과, 감마 혼합 모형이 검정력에 있어서 큰 이득을 볼 수 있는 것으로 나타났다.

주요용어: 연관 분석, 감마 혼합 모형, 항등 연결 함수, 다단계 일반화 선형 모형.

이 논문은 2014년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No.2013R1A1A1061332).

[1]교신저자: (402-751) 인천광역시 남구 용현동 235, 인하대학교 통계학과. E-mail: lwj221@gmail.com