*The Korean Journal of Applied Statistics* (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.2.295
**28**(2), 295–308

# Linear Mixed Models in Genetic Epidemiological Studies and Applications

Jeongmin Lim[a] · Sungho Won[b,1]

[a]Chunlab, Inc.; [b]Department of Public Health Science, Seoul National University

(Received March 23, 2015; Revised March 30, 2015; Accepted March 30, 2015)

**Abstract**

We have experienced a substantial improvement in and cost-drop for genotyping that enables genetic epidemiological studies with large-scale genetic data. Genome-wide association studies have identified more than ten thousand causal variants. Many statistical methods based on linear mixed models have been developed for various goals such as estimating heritability and identifying disease susceptibility locus. Empirical results also repeatedly stress the importance of linear mixed models. Therefore, we review the statistical methods related with to linear mixed models and illustrate the meaning of their estimates.

Keywords: Linear mixed model, genetic epidemiological studies, genome wide association study.

## 1. 서론

다양한 질병의 유전적 원인을 규명하는 유전역학연구(genetic epidemiological studies)는 유전자 정보없이 표현형(phenotype) 정보만을 활용하여 질병의 유전적 특성을 연구하는 질병기반 연구와, 표지유전자(genetic marker)와 관심 표현형 사이의 상관성(correlation)을 도출하는 유전자기반 연구로 구분할 수 있다. 표현형 기반 연구는 가족자료를 이용하며, 유전율(heritability) 및 유전분리성(segregation) 분석과 같이 질병의 유전적 성향을 파악하는 연구로써 유전자기반 연구의 사전 연구로 활용된다. 반면 유전자기반 연구는 가족자료 혹은 환자-대조(case-control) 자료를 이용하여 표지유전자(genetic marker)와 질병간의 상관 여부를 분석하는 연구로써, 관련분석(association analysis)과 연관분석(linkage analysis) 등이 이에 해당한다. 관련분석(association analysis)은 주로 단일염기다형성(single nucleotide polymorphism), 그리고 연관분석(linkage analysis)은 초위성체(microsatellite)가 표지 유전자(genetic marker)로 주로 이용된다. 단일염기다형성은 인간 유전체 상에 1Kb 간격으로 최소 하나 이상으로 존재하므로 초위성체에 비해 질병의 원인 유전자 규명을 위한 표지 유전자로서 더욱 유용하다. 또한 최근 단일염기다형성을 사용한 전장유전체분석이 활발히 진행되고 있으므로, 본 논문에서는 단일염기다형성 기반 분석 방법론을 위주로 논할 것이다.

선형혼합모형은 유전역학연구의 분석 방법으로 가장 활발히 활용되고 있는 방법 가운데 하나이다. 예를 들어 단일염기다형성을 활용한 전장유전체관련분석(genome-wide association studies)은 지난 십

This study was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2013R1A1A2010437).
[1]Corresponding author: Department of Public Health Science, Seoul National University, Gwanak-ro, Gwanak-Gu, Seoul 151-742, Korea. E-mail: won1@snu.ac.kr

여 년간 가장 활발히 진행된 유전역학연구로써, 2005년 시력 감퇴와 연관된 H 보체 인자(complement factor; Klein 등, 2005)의 원인유전자 규명에 처음 활용된 이후, 몸무게를 비롯한 650여개의 표현형에 대하여 14000여개의 원인유전자를 규명하였다 (Welter 등, 2014). 전장유전체관련분석은 백만여 개의 단일염기다형성들이 표현형에 미치는 영향을 분석한다. 그러나 일반적으로 단일염기다형성의 수가 샘플의 크기보다 월등히 크기 때문에, 특정 단일염기다형성의 효과와 나머지 모든 단일염기다형성들의 효과 총합을 각각 고정효과(fixed effect)와 랜덤효과(random effect)로 포함하는 선형혼합모형이 주로 활용된다. 다유전자효과모형(polygenic effect model)에 따르면, 개별 유전자들의 표현형에 대한 효과가 작고 유전자간에 상호작용(interaction)이 없는 경우 다유전자효과(polygenic effect)는 정규분포를 따른다고 알려져 있다 (George와 Mcculloch, 1993). 또한 멘델의 법칙에 의하여 가족구성원들 간의 표현형은 서로 유사한 경향이 있다. 따라서 가족자료를 이용한 유전역학 분석도 전장유전체 자료분석과 마찬가지로, 다유전자효과를 랜덤효과로 모형화하는 선형혼합모형을 주로 활용한다. 이처럼 선형혼합모형은 유전역학분석에서 다방면에 활용되고 있다.

본 논문에서는 이처럼 유전역학분석에서 다양하게 활용되고 있는 선형혼합모형 관련 이론과 다양한 활용 사례를 소개하고자 한다. 논문의 구성은 다음과 같다. 우선 유전역학연구에서 다양하게 활용되고 있는 선형혼합모형을 제시한 후 표현형기반 연구로 다유전자효과 모형 및 가족자료를 활용한 유전율 추정 방법에 대하여 설명하였다. 또한 유전자기반 연구로써 독립자료 및 가족자료를 활용하는 전장유전체 분석과 전장유전체 단일염기다형성들의 통합효과의 의미와 추정 알고리즘을 설명하였다. 각 단원에서는 활용된 선형혼합모형과 관련 소프트웨어를 소개하였다.

## 2. 가능도함수

### 2.1. 선형혼합모형

선형결합모형은 고정효과와 랜덤효과를 모두 포함하고 있는 모형으로, 보통 반복측정자료와 같이 반응변수 사이에 상관성이 있는 자료 분석에 주로 활용된다. 유전역학분석은 많은 경우 개체 간의 상관성이 존재하기 때문에 선형혼합모형이 광범위하게 사용되어 왔다. 현재까지 알려진 유전역학 선형혼합모형 기반 자료 분석방법 중에 Zhou와 Stephen에 의하여 제안된 GEMMA (Zhou와 Stephens, 2012) 알고리즘이 가장 빠른 방법이므로 이를 소개하고자 한다. 표현형(phenotype) 벡터 $\mathbf{y}$, 단일염기다형성의 수를 $M$, 그리고 절편(intercept)과 환경변수가 $P$개 있다고 하자. 이 때, 설계행렬(design matrix)을 $\mathbf{X}$라고 하면 $\mathbf{y}$의 차원은 $N\times 1$, $\mathbf{X}$는 절편, 환경변수 그리고 분석하고자 하는 한 개의 단일염기다형성으로 이루어진 $N\times(P+1)$이다. 유전적 요인에 대한 분산을 $\sigma_g^2$, 환경적 요인에 의한 분산을 $\sigma^2$라고 하자. $\mathbf{I}_w$는 $w\times w$ 정방행렬이라고 정의하면, 일반적으로 다음의 선형혼합모형을 고려한다:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{b} + \varepsilon, \quad \mathbf{b} \sim \mathrm{MVN}\left(0, \sigma_g^2\mathbf{A}\right), \ \varepsilon \sim \mathrm{MVN}\left(0, \sigma^2\mathbf{I}_N\right), \tag{2.1}$$

여기서 $\mathbf{A}$는 사전에 모든 원소가 알려진 상수행렬로써, Kinship Coefficient(KC) 행렬, Identity-By-Descent(IBS) 행렬 혹은 Genetic Relationship(GR) 행렬 등이 사용될 수 있다. 각 행렬은 다음 단원에서 자세히 설명할 것이다.

### 2.2. 최대가능도(Maximum likelihood) 추정량

$\lambda = \sigma_g^2/\sigma^2$이고 $\mathbf{H} = \lambda\mathbf{A} + \mathbf{I}$라고 가정하면, 로그가능도함수는 다음과 같다 (Zhou와 Stephens, 2012):

$$\log L\left(\lambda, \sigma^2, \beta\right) = -\frac{N}{2}\log\sigma^2 - \frac{N}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{H}| - \frac{1}{2}\sigma^{-2}(\mathbf{y}-\mathbf{X}\beta)^t\mathbf{H}^{-1}(\mathbf{y}-\mathbf{X}\beta). \tag{2.2}$$

따라서 $\mathbf{P}_X = \mathbf{H}^{-1} - \mathbf{H}^{-1}\mathbf{X}(\mathbf{X}^t\mathbf{H}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{H}^{-1}$이라고 하면 $\lambda$의 profile 가능도함수(profile likelihood function), 점수함수(score function)와 정보함수(information function)는 다음과 같다:

$$\begin{aligned}
l(\lambda) &= -\frac{N}{2}\log\frac{N}{2\pi} - \frac{N}{2}\log|\mathbf{H}| - \frac{N}{2}\log\left(\mathbf{y}^t\mathbf{P}_X\mathbf{y}\right), \\
\frac{\partial \log L}{\partial \lambda} &= -\frac{1}{2}\mathrm{tr}\left(\mathbf{H}^{-1}\mathbf{A}\right) + \frac{N}{2}\cdot\frac{\mathbf{y}^t\mathbf{P}_X\mathbf{\Phi}\mathbf{P}_X\mathbf{y}}{\mathbf{y}^t\mathbf{P}_X\mathbf{y}}, \\
\frac{\partial^2 \log L}{\partial \lambda^2} &= \frac{1}{2}\mathrm{tr}\left(\mathbf{H}^{-1}\mathbf{A}\mathbf{H}^{-1}\mathbf{A}\right) - \frac{N^2}{2}\cdot\frac{\left(\mathbf{y}^t\mathbf{P}_X\mathbf{A}\mathbf{P}_X\mathbf{A}\mathbf{P}_X\mathbf{y}\right)\left(\mathbf{y}^t\mathbf{P}_X\mathbf{y}\right) - \left(\mathbf{y}^t\mathbf{P}_X\mathbf{A}\mathbf{P}_X\mathbf{y}\right)^2}{\left(\mathbf{y}^t\mathbf{P}_X\mathbf{y}\right)^2}.
\end{aligned}$$

선형혼합모형에서 사용되는 $\mathbf{A}$는 모든 단일염기다형성 자료에 동일하게 적용되므로 고유값 분해(eigenvalue decomposition)를 활용하여 계산량을 줄일 수 있다. $\mathbf{A}$가 다음과 같이 표현된다고 가정하자:

$$\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^t. \tag{2.3}$$

위 식에서 $\mathbf{D}$는 고유치가 크기순으로 정렬된 대각행렬이고, $\mathbf{P}$와 $\mathbf{P}^t$는 $\mathbf{A}$의 고유벡터로 이루어진 행렬로써, 전치행렬이다. 만약 $\mathbf{y}$, $\mathbf{X}$를 각각 $\mathbf{y}_P = \mathbf{P}^t\mathbf{y}$, $\mathbf{X}_P = \mathbf{P}^t\mathbf{X}$로 변환하고 정보함수를 이용하여 추정된 $\lambda$의 추정량을 $\hat{\lambda}$이라고 하면, $(\sigma^2, \beta)$의 추정식은 $\mathbf{y}_P$, $\mathbf{X}_P$를 활용하여 나타낼 수 있다:

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{N}\mathbf{y}_P^t\mathbf{P}^t\mathbf{P}_X\mathbf{P}\mathbf{y}_P, \\
\hat{\beta} &= \left(\mathbf{X}_P^t\left(\hat{\lambda}\mathbf{D} + \mathbf{I}_N\right)^{-1}\mathbf{X}_P\right)^{-1}\mathbf{X}_P^t\left(\hat{\lambda}\mathbf{D} + \mathbf{I}_N\right)^{-1}\mathbf{y}_P.
\end{aligned} \tag{2.4}$$

따라서 선형혼합모형식 (2.1)의 모수 추정 알고리즘의 복잡도는 $\lambda$의 추정알고리즘에 영향을 받음을 알 수 있다. $\lambda$를 추정하기 위한 다양한 알고리즘이 제안되어 왔다. 만약 $M$개의 서로 다른 단일염기다형성 자료에 선형혼합모형식 (2.1)을 $M$번 적합해야하는 경우 GEMMA알고리즘 (Zhou와 Stephens, 2012)이, 그리고 유전율의 추정과 같이 선형혼합모형식 (2.1)을 1번만 적합하는 경우 평균 정보(average information)를 활용한 방법 (Gilmour 등, 1995)의 모수 추정 시간이 가장 빠르다고 알려져 있다.

### 2.3. 제한최대가능도(Restricted Maximum Likelihood) 추정량

최대가능도 방법은 분산 모수 추정시 평균모수의 추정량을 활용하기 때문에 모수의 추정치가 가지고 있는 편의가 커질 수 있다 (Smyth와 Verbyla, 1996). 따라서 선형혼합모형을 이용하는 경우 분산 모수는 제한가능도 방법을 이용하여 추정한다 (Corbeil와 Searle, 1976; Kenward와 Roger, 1997). 분산 모수 $(\sigma^2, \lambda)$ 추정을 위한 제한가능도함수는 다음과 같다:

$$\begin{aligned}
\log L\left(\sigma^2, \lambda\right) = &-\frac{N-P-1}{2}\log\sigma^2 - \frac{N-P-1}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{H}| \\
&- \frac{1}{2\sigma^2}\mathbf{y}^t\mathbf{P}_X\mathbf{y} - \frac{1}{2}\log\left|\mathbf{X}^t\mathbf{H}^{-1}\mathbf{X}\right|.
\end{aligned} \tag{2.5}$$

$\sigma^2$은 쉽게 추정이 가능하므로, $\lambda$의 추정을 위한 로그제한가능도함수를 구하고, 점수함수와 정보함수를 다음과 같이 구할 수 있다:

$$\begin{aligned}
\frac{\partial \log L}{\partial \lambda} &= -\frac{1}{2}\mathrm{tr}(\mathbf{P}_X\mathbf{A}) + \frac{N-P-1}{2}\cdot\frac{\mathbf{y}^t\mathbf{P}_X\mathbf{A}\mathbf{P}_X\mathbf{y}}{\mathbf{y}^t\mathbf{P}_X\mathbf{y}}, \\
\frac{\partial^2 \log L}{\partial \lambda^2} &= \frac{1}{2}\mathrm{tr}(\mathbf{P}_X\mathbf{A}\mathbf{P}_X\mathbf{A}) - \frac{N-P-1}{2}\cdot\frac{2\left(\mathbf{y}^t\mathbf{P}_X\mathbf{A}\mathbf{P}_X\mathbf{A}\mathbf{P}_X\mathbf{y}\right)\left(\mathbf{y}^t\mathbf{P}_X\mathbf{y}\right) - \left(\mathbf{y}^t\mathbf{P}_X\mathbf{A}\mathbf{P}_X\mathbf{y}\right)^2}{\left(\mathbf{y}^t\mathbf{P}_X\mathbf{y}\right)^2}.
\end{aligned} \tag{2.6}$$

최대가능도방법과 마찬가지로 위 점수함수와 정보함수를 이용하여 추정된 $\lambda$의 추정량을 $\hat{\lambda}$이라고 하면, $(\sigma^2, \beta)$의 추정식은 다음과 같다:

$$\hat{\sigma}^2 = \frac{1}{N-P-1}\mathbf{y}_P^t\left(\mathbf{P}^t\mathbf{P}_X\mathbf{P}\right)\mathbf{y}_P, \tag{2.7}$$
$$\hat{\beta} = \left(\mathbf{X}_P^t\left(\hat{\lambda}\mathbf{D}+\mathbf{I}_N\right)^{-1}\mathbf{X}_P\right)^{-1}\mathbf{X}_P^t\left(\hat{\lambda}\mathbf{D}+\mathbf{I}_N\right)^{-1}\mathbf{y}_P.$$

## 3. 선형혼합모형의 활용

### 3.1. 표현형기반 연구

**3.1.1. 다유전자효과(polygenic-effect) 모형** 키, 몸무게를 비롯한 대부분의 표현형들의 경우 개별 유전자들의 효과 크기는 미미하나, 다수의 원인 유전자들이 표현형에 영향을 미치고 있음이 알려져 있다. 다유전자효과모형이란 이처럼 개별 유전자의 효과크기는 작으나 다수의 유전자들이 표현형에 영향을 미치는 경우를 가정한다. 다유전자효과모형에 따르면, 각 개체의 다유전자효과는 정규분포를 따르며 멘델의 법칙에 의하여 가족 구성원들의 표현형은 유사한 경향이 있다. $n$개의 가족이 있고 $i$번째 가족의 구성원 수를 $n_i$라고 하면, $N = \sum n_i$이라고 할 수 있다. 각 개체의 상동염색체는 한 쌍으로 이루어져 있으므로, $i$번째 가족의 $j$번째 구성원의 한 쌍의 염색체의 다유전자효과를 각각 $g_{ij1}$, $g_{ij2}$라고 가정하자. 만약 $E(g_{ij1}) = E(g_{ij2}) = 0$, $\mathrm{var}(g_{ij1}) = \mathrm{var}(g_{ij2}) = \sigma_g^2$이고 $g_{ij1}, g_{ij2}$가 같은 분포를 따른다고 가정하자. 그리고 $i$번째 가족의 $j$번째 구성원의 총 다유전자효과를 $g_{ij}$라고 하고, 유전자간(interlocus)에 상호작용은 존재하지 않는다고 가정하자. $\mathrm{var}(g_{ij1}) + \mathrm{var}(g_{ij2})$은 다유전자가법효과분산(polygenic additive effect variance)라고 하고 $\sigma_a^2$으로 표기하자. 이는 유전자내(intralocus) 상호작용이 존재하지 않는 경우 즉, $g_{ij} = g_{ij1} + g_{ij2}$일 때 $g_{ij}$의 분산과 같다. 만약 유전자내 상호작용이 존재하는 경우 $\mathrm{var}(g_{ij})$는 $\sigma_a^2$보다 크며 그 차이를 $\sigma_d^2$라고 하자. 이때 $\sigma_d^2$는 다유전자우성효과분산(polygenic dominant effect variance)라고 한다. 만약 유전자내 상호작용이 존재하지 않는다면, $j$, $j'$의 다유전자효과들의 공분산은 아래와 같다 (Falconer, 1989):

$$\begin{aligned}\mathrm{cov}\left(g_{ij}, g_{ij'}\right) &= \mathrm{cov}\left(g_{ij1}+g_{ij2}, g_{ij'1}+g_{ij'2}\right) = E\left(g_{ij1}g_{ij'1}+g_{ij1}g_{ij'2}+g_{ij2}g_{ij'1}+g_{ij2}g_{ij'2}\right)\\ &= P(g_{ij1}=g_{ij'1})\mathrm{var}(g_{ij1}) + P(g_{ij1}=g_{ij'2})\mathrm{var}(g_{ij1})\\ &\quad + P(g_{ij2}=g_{ij'1})\mathrm{var}(g_{ij2}) + P(g_{ij_2}=g_{ij'2})\mathrm{var}(g_{ij2})\\ &= 2P(g_1=g_{ij'1})(2\mathrm{var}(g_{ij1})) = 2P(g_1=g_{ij'1})\sigma_a^2.\end{aligned}$$

이때 $P(g_{ij1} = g_{ij'1})$은 $i$와 $i'$ 사이의 혈연계수(kinship coefficient)라고 하고, 앞으로 $P(g_{ij1}{=}g_{ij'1})$는 $\pi_{ij,ij'}$으로 표기할 것이다. $\pi_{ij,ij'}$는 서로 다른 두 대립유전자가 identical-by-descent(IBD)일 확률을 이용하여 계산할 수 있다. 서로 다른 두 대립유전자가 IBD 관계에 있다는 의미는 두 대립유전자가 과거 동일 조상의 대립유전자의 복제본(replicate)임을 나타낸다. 만약 서로 다른 두 사람의 IBD 관계의 대립유전자가 2일 확률을 $f_2$라고 하고, IBD 관계의 대립유전자가 1일 확률을 $f_1$, 그리고 IBD 관계의 대립유전자가 0일 확률을 $f_0$라 하자. 예를 들어, 일란성 쌍둥이끼리의 $(f_2, f_1, f_0)$은 $(1,0,0)$이며, 부모와 자식 간에는 $(0,1,0)$이다. 이때, $\pi_{ij,ij'} = 0.25f_1 + 0.5f_2$이고, 따라서 Table 3.1을 얻을 수 있다.

$i$번째 가족에 대한 KC 행렬을 $\mathbf{\Phi}_i$이라고 하면 $(\mathbf{\Phi}_i)_{j,j'}{=}2{\times}\pi_{ijij'}$이고, 근친결혼이 없는 경우 대각 원소는 1이다. 가족 구성원 간의 표현형의 유사성은 다유전자효과에 의해 발생한다고 가정하면, $i$번째 가족에 대하여 다음의 모형을 가정할 수 있다 (Falconer, 1989):

$$y_i = X_i\beta + g_i + \varepsilon_i, \quad g_i \sim \mathrm{MVN}\left(0, \sigma_a^2\mathbf{\Phi}_i\right),\ \varepsilon_i \sim \mathrm{MVN}\left(0, \sigma^2\mathbf{I}_{n_i}\right). \tag{3.1}$$

**Table 3.1.** The kinship coefficients for several common relationships.

| Relationship | $f_2$ | $f_1$ | $f_0$ | Kinship Coefficient |
|---|---|---|---|---|
| Monozygotic twin | 1 | 0 | 0 | 1/2 |
| parent-offspring | 0 | 1 | 0 | 1/4 |
| Siblings | 1/4 | 1/2 | 1/4 | 1/4 |
| Grandparent-grandchild | 0 | 1/2 | 1/2 | 1/8 |
| First cousin | 0 | 1/4 | 3/4 | 1/16 |

만약 유전자내 상호작용이 존재하는 경우, $\mathbf{\Psi}_i$는 $n_i \times n_i$ 정방행렬이고, $\mathbf{\Psi}_i$의 $(j, j')$원소는 $j, j'$ 사이의 $f_2$라고 가정하면, 다음의 선형혼합모형을 얻을 수 있다 (Falconer, 1989):

$$y_i = X_i\beta + g_i + \varepsilon_i, \quad g_i \sim \text{MVN}\left(0, \sigma_a^2\mathbf{\Phi}_i + \sigma_d^2\mathbf{\Psi}_i\right), \ \varepsilon_i \sim \text{MVN}\left(0, \sigma^2\mathbf{I}_{n_i}\right). \tag{3.2}$$

위 모형에서 $\sigma_d^2$는 $f_2 > 0$인 가족 구성원 즉, 형제 혹은 일란성쌍둥이와 같은 겹선형 관계(bilinear relationship)를 갖는 가족구성원이 데이터에 포함된 경우에만 추정가능하며, Lim 등 (2014)이 제안한 알고리즘을 이용하여 분산모수를 추정할 수 있다.

**3.1.2. 가족자료를 활용한 유전율 추정** 유전율이란 표현형에 영향을 미치는 유전적, 비유전적 효과 가운데 유전적 효과의 비율을 의미한다 (Falconer, 1989). 만약 가족 구성원들의 표현형의 유사성은 유전적 요인으로 인하여 발생한다고 가정하면, 유전율은 가족 구성원들의 표현형의 공분산을 이용하여 추정할 수 있다. 즉, 다유전자 사이에 상호작용은 존재하지 않는다고 가정하면, 다유전자효과모형 (3.1)에 의하여 다유전자 전체효과의 분산은 $\sigma_g^2 = \sigma_a^2 + \sigma_d^2$이다. 이때 협의(narrow-sense), 광의(broad-sense) 유전율 ${h_n}^2$, ${h_b}^2$은 각각 다음과 같이 정의된다 (Falconer, 1989):

$$h_n^2 = \frac{\sigma_a^2}{\sigma_g^2 + \sigma^2}, \qquad h_b^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma^2}. \tag{3.3}$$

그러나 유전율의 추정에서 가정하는 다유전자효과모형 (3.1), (3.2)는 다음의 비현실적인 가정을 하고 있다. 첫째, 환경적 영향으로 인하여 발생하는 표현형의 유사성은 존재하지 않는다고 가정한다. 그러나 환경적 요인으로 인한 공분산은 모수화(parameterization)가 쉽지 않아 많은 경우 모형에서 제외된다. 결과적으로 $\sigma_a^2$가 팽창(inflation)되어 유전율 추정값은 실제보다 크게 나오는 경향이 있다 (Manolio 등, 2009). 둘째, 다유전자들 사이에 상호작용은 존재할 것으로 예측되나 다유전자효과모형은 상호작용이 존재하지 않는다고 가정한다. Zuk 등 (2012)은 다유전자효과모형에서 유전자간의 상호작용을 무시하는 경우, 이로 인한 분산은 $\sigma_a^2$의 크기를 증가시킴을 보였다. 마지막으로 환경과 유전자 사이의 교호작용이 존재할 수 있다. 환경변수들은 분석하고자 하는 데이터에 따라 다를 수 있고, 이 때문에 추정된 유전율은 분석에 활용한 데이터의 특성에 따라 다른 값을 가질 수 있다.

유전율은 다유전자효과 모형 (3.1), (3.2)를 이용하여 추정하며, 유전율 추정 소프트웨어에는 S.A.G.E. (Elston과 Gray-McGuire, 2004), SOLAR (Almasy와 Blangero, 1998) 등이 있다. S.A.G.E, SOLAR는 모두 가족 자료를 이용하여 $\sigma_a^2$, $\sigma_d^2$를 추정할 수 있다. S.A.G.E.는 일란성쌍둥이가 존재하는 경우 분산모수들의 추정이 불가능하고, 형제들 간의 공유되는 분산과 부부 간에 공유되는 분산의 추정은 가능하다. 부부 간에 공유되는 분산은 자연선택(natural selection), 양성선택(positive selection) 혹은 음성선택(negative selection)이 실제로 존재함을 보여준다. 선형모형은 다음과 같다:

$$y_i = X_i\beta + g_i + \varepsilon_i, \quad g_i \sim \text{MVN}\left(0, \sigma_a^2\mathbf{\Phi}_i + \sigma_d^2\mathbf{\Psi}_i\right), \ \varepsilon_i \sim \text{MVN}\left(0, \sigma^2\mathbf{\Omega}_i\right). \tag{3.4}$$

**Table 3.2.** Table of methods which can calculate Heritability.

| Methods | Languages | Estimable Variance component | | Estimator |
|---|---|---|---|---|
| | | $\sigma_a^2$ | $\sigma_g^2$ | (ML, REML) |
| S.A.G.E | C++ | O | O | ML |
| SOLAR | C++, FORTRAN | O | O | ML |
| POLY | C++ | O | O | ML |
| MX | LISREL, EQS | O | O | ML |
| GEMMA | C++ | O | X | ML, REML |

만약 $\mathbf{\Omega}_i$의 $j, j'$ 원소를 $(\mathbf{\Omega}_i)_{j,j'}$라고 하면 $(\mathbf{\Omega}_i)_{j,j'}$는 다음과 같이 정의된다:

$$(\Omega_i)_{j,j'} = \begin{cases} 1, & j = j', \\ \rho_1, & \text{siblings}, \\ \rho_2, & \text{spouses}, \\ 0, & \text{o.w.} \end{cases} \tag{3.5}$$

SOLAR는 쌍둥이 자료를 이용하여 유전율을 계산할 수 있으나, $\rho_1$ 및 $\rho_2$의 추정이 불가능하다. SOLAR와 S.A.G.E.이외에 가족자료를 활용하여 유전율을 계산할 수 있는 소프트웨어는 POLY (Chen과 Abecasis, 2006), MX (Posthuma와 Boomsma, 2005) 등이 존재한다.

Table 3.2는 유전율 추정에 활용할 수 있는 소프트웨어와 각 소프트웨어의 특징을 보여준다. 다유전자 효과모형 (3.1), (3.2)는 다변량 자료에 쉽게 확장될 수 있으며, 다변량 자료를 활용한 유전율 추정이 가능하다. 예를 들어 혈압과 관련된 유전율은 수축기혈압과 이완기혈압을 모두 활용하여 추정하는 경우 상대적으로 작은 샘플로 정확한 추정을 쉽게 생각할 수 있다. 만약 $q$개의 표현형의 유전율을 추정하는 경우, $y_{ij}^k$를 가족 $i$의 구성원 $j$의 표현형 $k$라고 하자:

$$y_i^k = \begin{pmatrix} y_{i1}^k \\ \vdots \\ y_{in_i}^k \end{pmatrix}, \qquad y_i^* = \begin{pmatrix} y_i^1 \\ \vdots \\ y_i^q \end{pmatrix}. \tag{3.6}$$

또한 이에 대응되는 설계행렬은 $X_i^*$이라고 하고 다유전자효과를 다음과 같이 정의하자:

$$g_i^k = \begin{pmatrix} g_{i1}^k \\ \vdots \\ g_{in_i}^k \end{pmatrix}, \qquad g_i^* = \begin{pmatrix} g_i^1 \\ \vdots \\ g_i^q \end{pmatrix}. \tag{3.7}$$

$\mathbf{V}_g$는 $q$개의 표현형들의 다유전자가법효과들간의 분산-공분산 행렬, $\mathbf{V}_e$는 $q$개 표현형들의 오차항의 분산-공분산행렬이라고 하자. 그리고 모든 표현형들의 다유전자우성효과분산($\sigma_d^2$)은 0이라는 가정하에서 $g_i^*$의 분산-공분산 행렬은 다음과 같이 정의할 수 있다:

$$\text{var}(g_i^*) = \mathbf{V}_g \otimes \mathbf{\Phi}_i. \tag{3.8}$$

따라서 다음의 선형혼합모형을 활용해 다변량 자료기반 유전율을 추정할 수 있다 (Vattikuti 등, 2012):

$$Y_i = \mathbf{X}_i^* \beta + \mathbf{g}_i^* + \varepsilon_i^*, \quad \mathbf{g}_i^* \sim \text{MVN}\left(\mathbf{0}, \mathbf{V}_g \otimes \mathbf{\Phi}_i\right), \ \varepsilon_i \sim \text{MVN}\left(\mathbf{0}, \mathbf{V}_\varepsilon \otimes \mathbf{I}_i\right). \tag{3.9}$$

이때 $k$번째 표현형에 대한 유전율은 $(\mathbf{V}_g)_{kk}/[(\mathbf{V}_g)_{kk} + (\mathbf{V}_\varepsilon)_{kk}]$이다. 다변량자료의 다유전자효과모형의 적합은 GEMMA (Zhou와 Stephens, 2012) 소프트웨어를 이용하거나, Vattikuti 등 (2012)에 의한 MINITAB 패키지를 이용하면 된다. 다변량 자료를 이용할 때 만약 적절한 분산-공분산을 구조를 이용하여 추정하는 경우 분산모수 추정량들의 표준오차는 작아지는 경향이 있음이 확인되었다 (Korte 등, 2012). 그러나 차원의 증가로 인한 계산량의 시간 복잡도를 줄이기 위한 연구가 필요하다.

### 3.2. 유전자기반 연구

**3.2.1. 독립자료를 이용한 전장유전체 분석** 전장유전체관련분석은 백만여 개에 이르는 단일염기다형성과 표현형간의 상관성을 검정함으로써 원인유전자를 규명한다. 그러나 서로 다른 인종 간에는 유전적 차이가 있기 때문에, 표현형 또한 인종별 차이가 있을 경우 인종 변수는 교락변수(confounding variable)가 될 수 있다. 이를 인구집단층화(population stratification) 라고 한다. 예를 들어, 식습관에 영향을 미치는 원인유전자를 규명한다고 가정하자. 식습관은 인종 별로 차이가 있기 때문에, 환자군과 대조군의 인종 분포가 다른 데이터를 이용하여 전장유전체관련 분석을 하면 지나치게 많은 거짓 양성(false positive) 결과를 얻을 수 있다. 또한 동일 인종임에도 거주지역에 따라 유전적 특성에 차이가 있을 수 있으므로 (Tang 등, 2005), 인구집단층화 문제에 로버스트한 분석 방법에 대한 연구가 진행되어 왔다 (Kang 등, 2008a, 2008b; Listgarten 등, 2010; Price 등, 2010; Yu 등, 2006; Zhang 등, 2010).

최근 인구집단층화 문제에 로버스트한 선형혼합모형이 제안되었다. 샘플 크기는 $N$, 전체 $M$개의 단일염기다형성이 있다고 가정하자. 또한 $M$개의 단일염기다형성들의 유전자 관측값으로 이루어진 $N \times M$ 설계행렬을 $\mathbf{G}$라고 하자. 독립샘플을 가정하므로 $n_1 = \cdots = n_n = 1$이다. 이 때 Kang 등 (2010)에 의하여 제안된 선형혼합모형은 환경변수와 분석하고자 하는 단일염기다형성은 고정변수로, 그리고 다유전자효과의 총합은 랜덤변수로 포함하며 선형모형은 다음과 같다:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{G}u + \varepsilon, \quad u \sim \text{MVN}\left(0, \sigma_g^2 \mathbf{I}_M\right), \ \varepsilon \sim \text{MVN}\left(0, \sigma^2 \mathbf{I}_N\right). \tag{3.10}$$

분석하고자 하는 하나의 단일염기다형성은 고정효과의 설계행렬 $\mathbf{X}$의 계수 벡터인 $\beta$에 포함된다. 결과적으로 모형 (3.10)를 $M$번 적합해야 된다. 모형 (3.10)는 다음 모형과 같이 표현할 수 있다:

$$\mathbf{y} = \mathbf{X}\beta + g + \varepsilon, \quad g \sim \text{MVN}\left(0, \sigma_g^2 \mathbf{G}\mathbf{G}^t\right), \ \varepsilon \sim \text{MVN}\left(0, \sigma^2 \mathbf{I}_N\right). \tag{3.11}$$

모형 (3.11)에 따르면 서로 다른 개체 사이의 유전적 차이는 랜덤효과의 분산-공분산 행렬인 $\mathbf{G}\mathbf{G}^t$에 의하여 보정됨을 알 수 있다. 결과적으로 각 개체 사이의 유전적 거리를 잘 반영할 수 있는 적절한 $\mathbf{G}\mathbf{G}^t$의 선택이 중요하다. Kang 등 (2010)은 IBS 행렬과 GR 행렬을 이용할 것을 제안하였다. $\text{ibs}(i, i', m)$은 단일염기다형성 $l$의 개체 $i$와 $i'$의 공통된 대립유전자 개수라고 하자. Table 3.3은 각 표지유전자의 유전형별 $\text{ibs}(i, i', m)$를 예를 들어 보여준다. 이때, IBS 행렬은 다음과 같이 정의한다:

$$\left(\mathbf{G}\mathbf{G}^t\right)_{ii'} = \frac{1}{M} \sum_{m=1}^{M} \frac{\text{ibs}(i, i', m)}{2}. \tag{3.12}$$

$x_{im}$는 단일염기다형성 $m$의 소수대립유전자(minor allele)의 수를 나타내는 값이고, $p_m$는 단일염기다형성 $m$의 소수대립유전자 빈도(minor allele frequency)라고 하자. 이때 GR 행렬은 다음과 같이 정의된다:

$$\left(\mathbf{G}\mathbf{G}^t\right)_{ii'} = \frac{1}{M} \sum_{m=1}^{M} \frac{(x_{im} - 2p_m)(x_{i'm} - 2p_m)}{2p_m(1 - p_m)}. \tag{3.13}$$

**Table 3.3.** Example of ibs $(i, i', m)$; Identity by state(IBS)

| Genotype of subject $i$ | Genotype of subject $i'$ | ibs$(i, i', m)$ |
|---|---|---|
| AA | AA | 2 |
| AA | Aa | 1 |
| AA | aa | 0 |
| Aa | AA | 1 |
| Aa | Aa | 2 |
| Aa | aa | 1 |
| aa | AA | 0 |
| Aa | Aa | 1 |
| Aa | aa | 2 |

표현형이 이진형인 경우 모형 (3.10), (3.11)을 활용할 수 없다. 일반화선형혼합모형(generalized linear mixed model)을 활용할 수 있으나, 가능도함수의 정확한 계산이 어렵고 모수가 수렴하지 않는 경우가 많아 자주 활용되지는 않는다 (Price 등, 2010). 대신 근사적인 방법으로 $\mathbf{GG}^t$의 고유치를 공변량으로 포함하여 분석한다. $\mathbf{GG}^t$을 고유치 분해한 결과가 다음과 같다고 가정하자 (Price 등, 2006):

$$\mathbf{GG}^t = \mathbf{PDP}^t, \quad \mathbf{P} = (e_1, \ldots, e_N). \tag{3.14}$$

만약 $K$개의 고유벡터를 모형에 포함시킨다고 가정하면, 다음의 로지스틱 회귀모형을 적합한다:

$$\log \frac{E(y)}{1 - E(y)} = \mathbf{X}\beta + \sum_{k=1}^{K} e_k \tau_k. \tag{3.15}$$

그러나 키와 몸무게 같이 유전율이 높은 표현형의 경우 모형에 포함한 고유벡터들의 분산 설명력이 충분히 크지 않기 때문에 분석 결과에 편의(bias)가 발생할 수 있다 (Zhou와 Stephens, 2012).

선형혼합모형을 활용한 전장유전체관련분석은 백만여 개의 단일염기다형성을 검정해야 하므로, 선형혼합모형을 백만여 번 적합해야 한다. 따라서 샘플 크기가 5000이상 되는 경우 계산 시간이 지나치게 많이 소요되며, 이러한 계산량의 문제를 해결하기 위한 다양한 알고리즘이 개발되었다 (Aulchenko 등, 2007b; Kang 등, 2008b; Zhou와 Stephens, 2012). Aulchenko 등 (2007b)에 의하여 개발된 GenABEL은 근사적인 방법을 활용하여 계산속도를 줄인다. 일반적으로 개별 단일염기다형성이 표현형에 미치는 효과크기는 작으므로, 단일염기다형성을 공변량으로 포함하지 않은 선형혼합모형을 적합하여 잔차를 추정하고, 잔차를 반응변수로 활용하는 분석방법을 제안하였다 (Aulchenko 등, 2007a). 잔차를 계산하기 위하여 이용하는 선형혼합모형은 설계행렬 $\mathbf{X}$에 단일염기다형성이 없다는 점을 제외하고 모형식 (3.11)와 동일하다. 모형식 (3.11)를 적합하여 얻은 잔차 벡터를 $\hat{\varepsilon}$라고 하면, $\hat{\varepsilon}$를 반응변수로 하고, 단일염기다형성 유전자형을 설명변수로 하는 단순회귀모형을 적합하면 된다.

그러나 GenABEL은 유전율이 크거나 효과크기가 큰 단일염기다형성이 존재하는 경우 결과에 심각한 편의가 발생할 수 있다. Kang 등 (2008b)은 고유치 분해 방법을 이용한 EMMA 알고리즘을 제안하였다. EMMA 방법은 GenABEL과 달리 각 단일염기다형성들의 효과크기와 분산모수를 정확하게 계산하는 방법이다. Lippert 등 (2011), Zhou와 Stephens (2012)은 EMMA의 계산속도를 더욱 향상시킨 FaST-LMM, GEMMA 알고리즘을 각각 제안하였다. $M$은 SNP의 수, $N$은 샘플의 크기, $P$는 공변량의 수라고 하자. 그리고 $t$는 분산 모수의 추정치가 수렴에 이르기 위해 필요한 반복의 수라고 하자. 이 때, Table 3.4 (Zhou와 Stephens, 2012)는 현재까지 제안된 방법별 계산 복잡도를 보여주며, 근사적으로 추정량을 계산하는 GenABEL을 제외했을 때 GEMMA 방법이 가장 계산속도가 빠름을 알 수 있다.

**Table 3.4.** Performance of different methods for GWAS with the Linear mixed model

| Methods | Time complexity |
|---|---|
| GEMMA (Zhou and Stephens, 2012) | $O(N^3+PN^2+MN^2+MTP^2N)$ |
| EMMA (Kang *et al.*, 2008b) | $O(N^3+MN^3+MTN)$ |
| FaST-LMM (Lippert *et al.*, 2011) | $O(N^3+PN^2+MN^2+MTP^2N)$ |
| GenABEL (Aulchenko *et al.*, 2007b) | $O(N^3+ MN^2+TN)$ |
| EMMAX (Kang *et al.*, 2010) | $O(N^3+ MN+TN)$ |

**Table 3.5.** Performance of different methods

| Methods | Computing time |
|---|---|
| GEMMA (Zhou and Stephens, 2012) | 3.3hours |
| EMMA (Kang *et al.*, 2008b) | 27years |
| FaST-LMM (Lippert *et al.*, 2011) | 6.2hours |
| GenABEL (Aulchenko *et al.*, 2007b) | 12mininutes |
| EMMAX (Kang *et al.*, 2010) | 6.4hours |

GEMMA 방법은 모수 추정에 있어서 행렬의 연산이 포함되어 있는 가능도 함수, 점수함수, 정보함수를 각 단일염기다형성마다 EMMA 방법보다 쉽고 효율적으로 계산하여 추정 알고리즘에서 사용하기 때문에 계산속도가 가장 빠르다. 또한 Table 3.5 (Zhou와 Stephens, 2012)는 샘플의 크기가 5000명, SNP의 수가 442,001개일 때 각 방법별 계산 시간을 보여준다:

**3.2.2. 가족자료를 이용한 전장유전체 분석** 초창기 전장유전체 분석은 가족자료를 기반으로 연관분석을 활용하여 멘델리안 표현형의 원인 유전자의 위치를 파악하였다 (Ott, 1999; Ott 등, 1974). 그러나 연관분석은 대가족 자료를 필요로 하고, 단일염기다형성 표지 유전자에 비하여 원인유전자의 위치를 파악하는데 있어 상대적으로 부정확한 위치 정보를 주기 때문에 최근 연관연구보다 관련연구가 활성화되었다 (Risch와 Merikangas, 1996). $n$개의 가족, $i$번째 가족의 구성원의 수가 $n_i$라고 하자. 단일염기다형성 자료를 이용한 가족자료 기반 전장유전체분석은 유전율 추정을 위한 선형혼합모형 (3.2)과 동일해야 한다. 다만 설계행렬은 단일염기다형성을 포함해야 한다.

가족 자료를 이용하여 전장유전체 분석을 할 수 있는 소프트웨어로는 SOLAR, S.A.G.E가 있다. Table 3.6는 선형혼합모형 이외의 기타 다른 모형을 활용하여 가족 자료를 이용한 관련분석을 할 수 있는 소프트웨어를 보여준다 (Ott 등, 2011). 또한 만약 $\sigma_d^2 = 0$을 가정할 수 있는 경우, 독립자료를 활용한 전장유전체관련분석에 활용할 수 있는 모든 소프트웨어를 원칙적으로 활용할 수 있다.

**3.2.3. 전장유전체 통합효과 추정** 전장유전체관련분석의 경우 수백만여 개에 이르는 단일염기다형성을 검정하기 때문에 다중비교문제로 인하여 보통 $10^{-7}$을 유의수준으로 사용한다. 또한 대부분의 원인유전자들의 효과크기는 작기 때문에, 통계분석의 검정력은 굉장히 낮은 편이다. 따라서 원인유전자를 찾기 위한 분석 이외에 전체 전장유전체 단일염기다형성 자료에 의하여 설명되는 표현형 분산의 비율을 계산함으로써, 각 표현형 별로 단일염기다형성-칩 자료의 상대적 중요성을 정량화하기 위한 시도가 이루어졌다 (Yang 등, 2011). 이렇게 계산된 비율이 유전율과 차이가 많이 나는 경우, 이는 단일염기다형성-칩 자료 외에 차세대염기서열 자료 기반 희소유전자와 같은 기타 다른 유전적 인자들을 활용하여 표현형간과의 연관성을 검정해야 할 필요가 있음을 의미한다. “설명되지 않는 유전율(missing heritability)”이라고 불리는 이 추정값은 (Hindorff 등, 2009) 가족데이터에서 추정한 유전율과 단일염기다형성-칩에 존재하는 표지유전자들에 의하여 설명되는 분산비율의 차를 의미한다. 예를 들어, 키의

**Table 3.6.** Software packages for family-based association analysis

| Program name | Purpose |
|---|---|
| APL (Martin *et al.*, 2003) | 연관성(linkage)이 존재하는 경우에 관련성(association) 분석 |
| FBAT (Rabinowitz and Laird, 2000) | 멘델의 법칙을 활용한 스코어통계량을 계산 |
| MERLIN (Abecasis *et al.*, 2002) | 가족 구성원간의 IBD 행렬을 추정하고, 이를 바탕으로 선형혼합모형 기반 Wald 통계량을 계산 |
| PLINK (Purcell *et al.*, 2007) | Transmission disequilibrium test기반 trio 자료분석 통계량을 계산 |

경우 유전율이 대략 80%이나 현재까지 알려진 원인단일염기다형성에 의하여 설명되는 유전율은 5%이다 (Yang 등, 2010).

단일염기다형성-칩 자료의 표지유전자들에 의하여 설명되는 표현형의 분산비율은 독립자료를 활용한 전장유전체관련분석 (3.11)과 유사한 선형혼합모형을 활용한다. 모형 (3.11)와 비교했을 때, 세 가지 측면에서 차이가 있다. 첫째 설계행렬 $\mathbf{X}$는 단일염기다형성자료를 포함하지 않는다. 둘째, $\mathbf{GG}^t$행렬로써 GR 행렬을 사용한다. 셋째, 모든 개체 사이에 $(\mathbf{GG}^t)_{ii'}$는 0.005보다 작아야 한다. 이 세 가정을 만족시키는 경우 모형 (3.11)에서 추정된 분산모수를 활용한 다음의 $h^2_{SNP}$는 단일염기다형성-칩에 의하여 설명되는 표현형 분산의 비율을 의미한다 (Yang 등, 2011):

$$h^2_{SNP} = \frac{M\sigma^2_g}{M\sigma^2_g + \sigma^2_\varepsilon}. \tag{3.16}$$

만약, 세 번째 가정이 만족되지 않는 경우 $h^2_{SNP}$은 협의의 유전율이 된다. $h^2_{SNP}$는 GCTA 소프트웨어 (Yang 등, 2011)를 활용하면 쉽게 계산할 수 있다.

또한 GCTA 소프트웨어는 기저 임계 모형(liability threshold model; Lee 등, 2011)을 활용하여 이진형 자료에 대한 단일염기다형성-칩 자료에 의하여 설명되는 분산의 비율을 계산하기 위하여 확장되었다. 질병의 유무는 연속형인 잠재변수에 의하여 결정되고, 잠재변수는 표준정규분포를 따른다고 가정하자 (Lynch와 Walsh, 1998). 만약 $\mathbf{l}$을 잠재변수들로 이루어진 벡터라고 가정하면, 다음의 선형혼합모형을 생각할 수 있다:

$$l = g + \varepsilon, \quad g \sim \text{MVN}\left(0, \sigma^2_g \mathbf{GG}^t\right), \ \varepsilon \sim \text{MVN}\left(0, \left(1-\sigma^2_g\right)\mathbf{I}_N\right). \tag{3.17}$$

위 모형식에서 추정된 분산 모수를 활용하여 다음을 정의할 수 있다:

$$\hat{h}^2_{\text{SNP}(L)} = \frac{M\hat{\sigma}^2_g}{M\hat{\sigma}^2_g + \hat{\sigma}^2_\varepsilon}. \tag{3.18}$$

이때 $\hat{h}^2_{SNP(L)}$은 이진형자료에서의 전체 단일염기다형성 자료에 의하여 설명되는 통합 효과 분산의 비율이라고 생각할 수 있다. 그러나 잠재변수 $\mathbf{l}$은 관찰되지 않으므로 $\hat{h}^2_{\text{SNP}(L)}$의 추정치를 얻기 위하여 추가 계산이 필요하다. 이진형자료를 0/1로 이루어진 연속형 반응변수로 간주하고 계산한 $\hat{h}^2_{\text{SNP}(L)}$을 $h^2_{\text{SNP}(O)}$이라고 하자. prev는 대상 모집단에서 표현형이 1인 개체의 비율(유병률)이며, $v$은 샘플 전체에서 질병에 걸린 케이스에 대한 비이고, $z$는 표준정규분포의 $1-\text{prev}$라고 하자. 이때 Lee 등 (2011)은 다음의 관계식을 도출하였다:

$$\hat{h}^2_{\text{SNP}(L)} = \hat{h}^2_{\text{SNP}(O)} \frac{\text{prev}(1-\text{prev})}{z^2} \frac{\text{prev}(1-\text{prev})}{v(1-v)}, \tag{3.19}$$

$$\text{s.e.}\left(\hat{h}^2_{\text{SNP}(L)}\right) = \text{s.e.}\left(\hat{h}^2_{\text{SNP}(O)}\right) \frac{\text{prev}(1-\text{prev})}{z^2} \frac{\text{prev}(1-\text{prev})}{v(1-v)}. \tag{3.20}$$

따라서 유병률에 대한 정보가 이용 가능한 경우, 환자-대조연구 자료에서 $\hat{h}^2_{SNP(L)}$의 추정이 가능함을 알 수 있다.

## 4. 결론

지난 반 세기 동안 유전형 기술(genotyping technology)의 발달로 인하여 적은 비용으로 많은 유전체 자료를 얻을 수 있는 시대가 도래하였다. 이로 인하여 인간 질병의 유전적 원인을 도출하기 위한 다양한 연구들이 활성화되었고, 동시에 다양한 유전체 분석 방법들이 개발되었다. 본 논문에서 정리한 것처럼 선형혼합모형은 전장유전체관련분석을 비롯하여 많은 유전역학연구에서 가장 빈번하게 활용되고 있는 모형으로, 앞으로도 다양한 유전체분석에 활용될 것으로 기대된다.

그러나 지난 수십 년 동안 다양한 분석 알고리즘들이 개발되어 왔음에도 불구하고, 아직까지 일부 분석 방법들에는 한계가 존재하고 추가 연구가 필요한 상황이다. 예를 들어, 표현형이 이진형인 경우, 각 구성원간의 상관성을 적절히 모형화하는 것은 아직도 어려운 상황이다. 가족 자료를 이용한 분석의 경우, 각 가족마다 구성원의 수가 다르고 분산-공분산 구조에 차이가 있기 때문에 일반화추정방정식(generalized estimating equations)을 활용하기 힘들고, 일반화선형혼합모형의 경우 모수 추정량이 계산되지 않는 경우가 빈번하여 각 구성원간의 상관성을 적절히 고려할 수 있는 분석 알고리즘의 개발이 필요하다. 또한 최근 차세대염기서열분석 기술 발달로 인하여, 유전체 자료의 규모는 날로 커지고 있어, 계산이 빠른 알고리즘의 필요성은 점차 증가되고 있다. 유전역학 분야에서의 선형혼합모형은 중요성을 고려할 때, 선형혼합모형은 앞으로도 다양하게 활용될 것으로 생각되며, 선형혼합모형의 적합에 소요되는 시간을 최소화할 수 있는 알고리즘의 개발은 앞으로도 지속적으로 요구될 것으로 기대한다.

## References


Abecasis, G. R., Cherny, S. S., Cookson, W. O. and Cardon, L. R. (2002). Merlin-rapid analysis of dense genetic maps using sparse gene flow trees, *Nature Genetics*, **30**, 97–101.

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

Aulchenko, Y. S., de Koning, D. J. and Haley, C. (2007a). Genomewide rapid association using mixed model and regression: A fast and simple method for genomewide pedigree-based quantitative trait loci association analysis, *Genetics*, **177**, 577–585.

Aulchenko, Y. S., Ripke, S., Isaacs, A. and Van Duijn, C. M. (2007b). GenABEL: An R library for genome-wide association analysis, *Bioinformatics*, **23**, 1294–1296.

Chen, W. M. and Abecasis, G. R. (2006). Estimating the power of variance component linkage analysis in large pedigrees, *Genet Epidemiol*, **30**, 471–484.

Corbeil, R. R. and Searle, S. R. (1976). Restricted Maximum Likelihood (REML) Estimation of Variance Components in Mixed Model, *Technometrics*, **18**, 31–38.

Elston, R. C. and Gray-McGuire, C. (2004). A review of the 'Statistical Analysis for Genetic Epidemiology' (S.A.G.E.) software package, *Hum Genomics*, **1**, 456–459.

Falconer, D. S. (1989). *Introduction to Quantitative Genetics*, (3rd ed.), Burnt Mill, Harlow, Essex, England.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling, *Journal of the American Statistical Association*, **88**, 881–889.

Gilmour, A. R., Thompson, R. and Cullis, B. R. (1995). Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models, *Biometrics*, **51**, 1440–1450.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S. and Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits, *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 9362–9367.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., Sabatti, C. and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies, *Nature Genetics*, **42**, 348-U110.

Kang, H. M., Ye, C. and Eskin, E. (2008a). Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots, *Genetics*, **180**, 1909–1925.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J. and Eskin, E. (2008b). Efficient control of population structure in model organism association mapping, *Genomics*, **178**, 1709–1723.

Kenward, M. G. and Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood, *Biometrics*, **53**, 983–997.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C. and Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration, *Science*, **308**, 385–389.

Korte, A., Vilhjalmsson, B. J., Segura, V., Platt, A., Long, Q. and Nordborg, M. (2012). A mixed-model approach for genome-wide association studies of correlated traits in structured populations, *Nature Genetics*, **44**, 1066−+.

Lee, S. H., Wray, N. R., Goddard, M. E. and Visscher, P. M. (2011). Estimating missing heritability for disease from genome-wide association studies, *American Journal of Human Genetics*, **88**, 294–305.

Lim, J., Sung, J. and Won, S. (2014). Efficient strategy for the genetic analysis of related samples with a linear mixed model, *Journal of the Korean Data and Information Science Society*, **25**, 1025–1038.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I. and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies, *Nature Methods*, **8**, 833-U894.

Listgarten, J., Kadie, C., Schadt, E. E. and Heckerman, D. (2010). Correction for hidden confounders in the genetic analysis of gene expression, *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 16465–16470.

Lynch, M. and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*, Sunderland, Mass.: Sinauer.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F., McCarroll, S. A. and Visscher, P. M. (2009). Finding the missing heritability of complex diseases, *Nature*, **461**, 747–753.

Martin, E. R., Bass, M. P., Hauser, E. R. and Kaplan, N. L. (2003). Accounting for linkage in family-based tests of association with missing parental genotypes, *American Journal of Human Genetics*, **73**, 1016–1026.

Ott, J. (1999). *Analysis of Human Genetic Linkage*, (3rd ed.), Baltimore: Johns Hopkins University Press.

Ott, J., Kamatani, Y. and Lathrop, M. (2011). Family-based designs for genome-wide association studies, *Nature Reviews Genetics*, **12**, 465–474.

Ott, J., Schrott, H. G., Goldstei, J. I., Hazzard, W. R., Allen, F. H., Falk, C. T. and Motulsky, A. G. (1974). Linkage studies in a large kindred with familial hypercholesterolemia, *American Journal of Human Genetics*, **26**, 598–603.

Posthuma, D. and Boomsma, D. I. (2005). Mx scripts library: Structural equation modeling scripts for twin and family data, *Behavior Genetics*, **35**, 499–505.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies, *Nature Genetics*, **38**, 904–909.

Price, A. L., Zaitlen, N. A., Reich, D. and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies, *Nature Reviews Genetics*, **11**, 459–463.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Sklar, P., de Bakker, P. I., Daly, M. J. and Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses, *American Journal of Human Genetics*, **81**, 559–575.

Rabinowitz, D. and Laird, N. (2000). A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information, *Human Heredity*, **50**, 211–223.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases, *Science*, **273**, 1516–1517.

Smyth, G. K. and Verbyla, A. P. (1996). A conditional likelihood approach to residual maximum likelihood estimation in generalized linear models, *Journal of the Royal Statistical Society Series B-Methodological*, **58**, 565–572.

Tang, H., Quertermous, T., Rodriguez, B., Kardia, S. L. R., Zhu, X. F., Brown, A., Pankow, J. S., Province, M. A., Hunt, S. C., Boerwinkle, E., Schork, N. J. and Risch, N. J. (2005). Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies, *American Journal of Human Genetics*, **76**, 268–275.

Vattikuti, S., Guo, J. and Chow, C. C. (2012). Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits, *Plos Genetics*, **8**.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L. and Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations, *Nucleic Acids Research*, **42**(D1), D1001–D1006.

Yang, J. A., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E. and Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height, *Nature Genetics*, **42**, 565-U131.

Yang, J. A., Lee, S. H., Goddard, M. E. and Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis, *American Journal of Human Genetics*, **88**, 76–82.

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S. and Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness, *Nature Genetics*, **38**, 203–208.

Zhang, Z. W., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M. and Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies, *Nature Genetics*, **42**, 355-U118.

Zhou, X. and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies, *Nature Genetics*, **44**, 821-U136.

Zuk, O., Hechter, E., Sunyaev, S. R. and Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability, *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 1193–1198.

# 선형혼합모형의 역할 및 활용사례: 유전역학 분석을 중심으로

임정민[a] · 원성호[b,1]

[a](주) 천랩, [b]서울대학교 보건대학원

(2015년 3월 23일 접수, 2015년 3월 30일 수정, 2015년 3월 30일 채택)

**요 약**

지난 수십 년 동안 유전형 기술(genotyping technology)의 발달로 개인별 유전자 정보를 얻기 위해 필요한 비용이 감소함에 따라, 다양한 인간 질병의 원인 유전자를 규명하기 위한 많은 유전역학 연구들이 진행되어 왔다. 예를 들어 전장유전체관련분석(genome-wide association studies)은 수백 개에 이르는 표현형(phenotypes)에 대하여 수천 개에 이르는 원인유전자를 규명하였다. 유전체 자료의 홍수로 인하여 대규모 유전체 자료를 분석할 수 있는 다양한 분석 알고리즘에 개발되었으며, 특별히 선형혼합모형은 유전율의 추정부터 관련분석(association studies)에 이르기까지 유전역학 연구에서 광범위하게 활용되고 방법론이었다. 본 논문에서는 유전역학 연구에 있어 빈번하게 활용되는 선형혼합모형의 활용 사례를 나열하고, 각 분석 모형 별 추정치들의 생물학적 의미를 논하고자 한다.

주요용어: 선형혼합모형, 유전역학연구, 전장유전체분석.

이 논문은 2013년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구 사업임 (2013R1A1A2010437).

[1]교신저자: (151-742) 서울특별시 관악구 관악로 1번지, 서울대학교 보건대학원 보건학과.
E-mail: won1@snu.ac.kr