Journal of the Korean Data & Information Science Society
2014, **25**(5), 1025–1038

http://dx.doi.org/10.7465/jkdi.2014.25.5.1025
한국데이터정보과학회지

선형혼합모형을 이용한 유전체 자료분석방안에 대한 연구[†]

임정민[1] · 성주헌[2] · 원성호[3]

[1](주)천랩 · [23]서울대학교 보건대학원

접수 2014년 6월 30일, 수정 2014년 7월 24일, 게재확정 2014년 8월 5일

요 약

가족 자료를 활용한 연속형 표현형의 전장유전체분석 (genome-wide association analysis)은 주로 선형혼합모형을 이용하며, 분산공분산행렬은 가족 구성원간의 유전적 거리를 고려하여 결정된다. 그러나 가족 구성원들의 표현형의 유사성은 유전적 요인과 환경적 요인에 의하여 발생함에도 불구하고, 표현형의 유사성은 단지 유전적 요인에 의해서 발생한다고 가정한다. 예를 들어 키의 경우 부부 사이에 양의 상관관계가 존재하나 유전적 요인만 고려하여 독립으로 가정한다. 선형혼합 모형에서 분산공분산 구조를 잘못 가정하는 경우, 검정통계량의 1종 혹은 2종의 오류를 적절히 관리할 수 없다. 본 논문에서는 다양한 유형의 분산공분산구조를 가정할 수 있는 선형혼합모형과 이를 기반으로 한 검정통계량을 제안하였다. 모의실험을 통하여 제안한 방법이 기존의 모형보다 통계적 검정력이 우수함을 확인하였다. 또한 체질량지수 (body mass index; BMI)의 전장유전체 분석에 적용하여 기존에 알려지지 않은 새로운 원인 유전자를 규명하였다.

주요용어: 뉴튼-라프슨 방법, 선형혼합모형, 전장유전체분석, 제한최대가능도, 평균 정보 방법.

1. 서 론

지난 10년간 genotyping 기술의 발달과 더불어 인간 질병들의 원인 유전자를 규명하기 위한 다양한 유전체 연구들이 진행되었다. 특별히 인간 유전체에 퍼져 있는 많은 단일염기다형성 (single nucleotide polymorphism; SNP)과 표현형의 연관성을 분석하여, 원인 유전자를 규명하는 분석을 전장유전체분석 (genome-wide association studies)이라고 한다. 전장유전체분석은 제 2형 당뇨 여부, 알츠하이머 여부, 체질량지수 (body mass index; BMI) 등 관심 표현형을 주로 반응변수로 사용하고, 독립변수는 단일염기다형성과 나이, 성별 등 반응변수에 영향을 미치는 환경 변수로 구성된다. 전장유전체 분석은 보통 100,000개 이상의 SNPs을 분석하므로, 다중비교문제가 존재하고, 이는 Dunn (1961)에 의해 제안된 본페로니 수정 방법을 이용하여 보정한다. 그러나 다중비교문제의 보정은 전장유전체 분석의 유의수준을 아주 작게 만들기 때문에, 통계적 검정력이 우수한 분석 방안에 대한 연구가 필요하다.가족 자료를 이용한 전장유전체분석은 가족 구성원간 표현형의 유사성을 고려하여 분석해야 한다. 특별히 분산공분산 구조는 가족 관계에 따라 달라진다. 예를 들어 Figure 1.1의 가족1 (Family 1)과 가족2 (Family 2)는

[†] 이 논문은 2013년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구 사업임 (2013R1A1A2010437).
[1] (151-742) 서울특별시 관악구 관악로 1번지, 서울대학교 유전공학연구소 105동 307호 Chunlab, Inc., 연구원.
[2] (151-742) 서울특별시 관악구 관악로 1번지, 서울대학교 보건대학원 보건학과, 교수.
[3] 교신저자: (151-742) 서울특별시 관악구 관악로 1번지, 서울대학교 보건대학원 보건학과, 조교수.
E-mail: won1@snu.ac.kr

서로 다른 분산공분산 구조를 갖고 있다. 따라서 대부분의 가족자료 기반 전장유전체 분석은 불균형구조 (unbalanced structure)를 갖고 있고, SAS, SPSS와 같은 통계소프트웨어는 이러한 불균형 자료를 적절히 다룰 수 없기 때문에 다양한 분석프로그램이 개발되었다. 예를 들어 가족 자료 기반 전장유전체 분석 소프트웨어로는 Kang 등 (2010)에 의하여 제안된 선형혼합모형기반 EMMAX, Zhou와 Stephens (2012)에 의해 제안된 GEMMA 등이 있다.

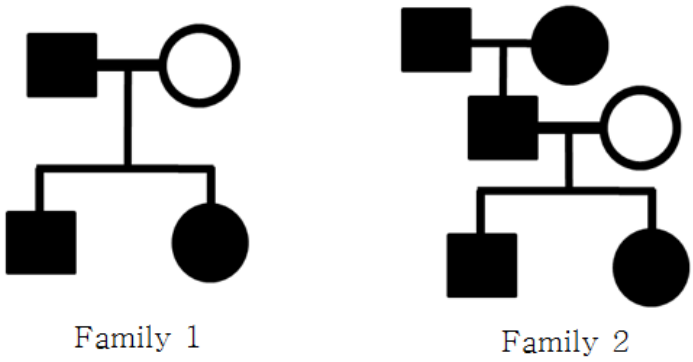

Figure 1.1 Unbalanced structure in family data

현재까지 개발된 가족자료 전장유전체 분석 소프트웨어는 가족 구성원간의 분산공분산 구조를 다유전자 효과 모형 (polygenic effect model; Valdar 등, 2006)을 활용하여 결정하였다. σ_g^2을 다유전자 효과의 분산이라고 가정하자. 가족 구성원의 반응변수의 유사성은 환경적 요인이 아닌 유전적 요인으로 생성되고, 다유전자의 효과가 가법적인 경우 임의의 두 가족 구성원간의 공분산은 KC (kinship coefficient)의 2 σ_g^2배임이 확인되었다 (Longman, 1996). KC는 멘델의 유전 법칙을 고려하여 계산되는 값으로 KC는 두 개체의 대립유전자를 무작위로 추출하였을 때, 두 대립유전자가 유전적으로의 동질적일 (identical by descent) 확률을 의미한다. 예를 들어 형제간의 KC는 1/4, 부부간의 KC는 0이다. Figure 1.1의 가족1, 2의 KC행렬은 다음과 같다:

$$\Phi_1 = \begin{bmatrix} 1 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix}, \quad \Phi_2 = \begin{bmatrix} 1 & 0 & \frac{1}{2} & 0 & \frac{1}{4} & \frac{1}{4} \\ 0 & 1 & \frac{1}{2} & 0 & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} & 1 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix}$$

따라서 가족 구성원간의 관계가 멀어질수록 KC도 작아지게 되고, 가족 구성원간의 표현형의 유사성을 적절히 고려할 수 있다.

그러나 다유전자 효과 모형은 유전적 요인에 의한 표현형의 유사성만을 고려할 뿐, 환경적 요인에 근거한 유사성은 존재하지 않는다고 가정한다. 따라서 본 논문에서는 가족 구성원 사이의 다양한 형태의 분산공분산 구조를 가정할 수 있는 선형혼합모형을 제안하였다. 모의실험 자료를 활용하여, 제안한 모형과 기존의 분석 방법의 경험적 검정력 (empirical power)을 비교하였다. 또한 제안한 모형을 HTK (healthy twin study, Korea) 코호트 자료를 활용한 BMI의 전장유전체분석에 적용하여, BMI의 새로운 원인유전자를 규명하였다.

본 논문은 총 4절로 구성되어있다. 1절에서는 전장유전체분석에 대한 설명과 기존 행해져 왔던 분석과 새로운 분석법을 제시하는 배경을 밝혔고, 2절에서는 새롭게 제안한 모형의 가능도함수와 모수 추정

방법을 제안하였다. 3절에서는 2절에서 언급한 방법을 모의실험 자료와 실제 데이터에 적용한 결과를 정리하였고, 마지막으로 4절에서는 제안한 방법의 장점 및 한계점을 최종적으로 논하였다.

2. 방법론

2.1. 가능도 함수

모수의 추정방법에는 Patterson과 Thompson (1971)이 제안한 최대가능도 (maximum likelihood; ML) 방법과 제한최대가능도 (restricted maximum likelihood; REML) 방법 두 가지가 있다. 두 방법은 모수의 수에 비하여 샘플의 크기가 상대적으로 큰 경우 근사적으로 비슷한 결과를 도출한다. 그러나 샘플의 크기가 작을 때, ML방법을 사용하면 추정치의 편의가 커질 수 있으므로 (Smyth와 Verbyla, 1996), 본 논문에서는 REML방법을 이용한 모수 추정 방법을 고려하였다. REML방법은 공변량의 회귀계수 β의 추정은 가능도 함수를 이용하고, 분산 모수의 추정은 제한가능도 (restricted likelihood; RL) 함수를 이용한다 (Searle, 1992).

반응변수의 벡터를 y, SNP와 환경변수로 구성된 행렬을 X라고 가정하자. 이때 환경변수는 절편을 포함한다. 개체의 수를 n환경변수의 수를 m이라고 하면 X는 $n \times (m+1)$차원 행렬이다. 또한 유전적 요인에 의한 분산을 σ_g^2, 환경적 요인에 의한 분산을 σ^2라고 표기하자. 가족구성원간의 공분산 구조가 환경적 요인과 유전적 요인에 의하여 발생하는 경우 유전적 요인에 의한 공분산은 $\sigma_g^2\Phi$, 환경적 요인에 의한 분산공분산행렬은 $\sigma^2 I + \sigma_1\Psi_1 + \cdots + \sigma_k\Psi_k$이라고 가정하자. 이때 Ψ_1,..., Ψ_k는 환경변수에 의하여 설명되는 공분산 구조를 고려하여 결정할 수 있다. 예를 들어 키와 같이 부부간의 양의 상관관계가 존재하는 경우 Figure 1.1에서 가족2의 Ψ_1는 다음과 같이 정의된다:

$$\Psi_1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

만약 형제간에 유사성이 추가적으로 존재하는 경우 다음과 같이 Ψ_2를 새롭게 정의해야 한다:

$$\Psi_2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

즉, $\Psi_1, \cdots, \Psi_k$은 기존의 KC행렬에서 설명하지 못한 환경적 요인을 설명하기 위하여 추가된 행렬이며, $\sigma_1, \cdots, \sigma_k$은 환경적 요인에 의하여 발생한 가족 구성원간의 공분산을 나타낸다. 개체의 수를 n, 환경요인의 수를 m라고 정의하면 결과적으로 다음의 선형혼합 모형을 고려할 수 있다:

$$y = X\beta + b + \epsilon, X \in R^{n\times(m+1)}$$

$$b \sim MVN(0, \sigma^2\lambda\Phi),\ \epsilon \sim MVN(0, \sigma^2(I_n + \rho_1\Psi_1 + \cdots + \rho_k\Psi_k)),\ \lambda = \sigma_g^2/\sigma^2,\ \rho_i = \sigma_i/\sigma^2$$

특별히 $H = \lambda\Phi + I_n + \rho_1\Psi_1 + \cdots + \rho_k\Psi_k$이라고 하면, y의 로그가능도함수 $l\,(\sigma^2, \lambda, \rho_1, \ldots, \rho_k, \beta)$는 다음과 같다:

$$l = -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi - \frac{1}{2}\log|H| - \frac{1}{2}(\sigma^2)^{-1}(y - X\beta)^t H^{-1}(y - X\beta)$$

만약 $(\sigma^2, \lambda, \rho_1, \cdots, \rho_k)$이 알려져 있다고 가정하면, $\widehat{\beta} = (X^t H^{-1} X)^{-1} X^t H^{-1} y$이다. 분산모수는 $A^t A = I_{n-m-1}$, $AA^t = I_n - X(X^t X)^{-1} X^t$을 만족하는 행렬 A에 대하여 $A^t y$의 가능도함수 즉, 제한가능도를 이용하여 추정하며 (Kenward와 Roger, 1997), 이때 로그제한가능도는 다음과 같다:

$$l_r = l + \frac{m+1}{2}\log\sigma^2 + \frac{m+1}{2}\log 2\pi - \log|X^t H^{-1} X|$$

2.2. 평균 정보 방법

모수의 초기값은 적률추정량 (method of moment estimator)을 이용하여 추정하였다. 가족 구성원 간에 독립을 가정할 때의 회귀계수의 추정량 $(X^t X)^{-1} X^t y$은 β의 불편성을 만족하므로, $e = y - (X^t X)^{-1} X^t y$를 이용하여 분산모수들의 초기값을 계산할 수 있다. vec는 벡터화 (vectorization) 연산자라고 (Neudecker와 Magnus, 1999) 정의하자. ϵ_e는 아래 회귀모형에서의 오차항 그리고 σ_e^2는 ϵ_e의 분산이라고 하고 하면, 다음의 모형을 가정하여 초기값을 추정할 수 있다:

$$vec(ee^t) = \sigma^2 vec(I) + \sigma_g^2 vec(\Phi) + \sum_{i=1}^{k} \sigma_i vec(\Psi_i) + \epsilon_e, \epsilon_e \sim N(0, \sigma_e^2)$$

다음으로 분산모수의 추정은 평균정보 (average information; AI) 방법을 이용하였다. 만약 $\theta = (\lambda, \rho_1, \ldots, \rho_k)^t$라고 하면, θ는 뉴튼-라프슨 (Newton-Raphson) 방법, 피셔의 점수화 (Fisher scoring) 방법, 그리고 이 둘을 혼합한 AI 방법 등을 이용하여 추정할 수 있다:

$$\text{뉴튼-라프슨} : \theta^{(n+1)} = \theta^{(n)} - \left(\frac{\partial^2 l}{\partial\theta\partial\theta^t}\right)^{-1} \left.\frac{\partial l}{\partial\theta}\right|_{\theta=\theta(n)}$$

$$\text{피셔의 점수화} : \theta^{(n+1)} = \theta^{(n)} - \left(E\left[\frac{\partial^2 l}{\partial\theta\partial\theta^t}\right]\right)^{-1} \left.\frac{\partial l}{\partial\theta}\right|_{\theta=\theta(n)}$$

$$AI : \theta^{(n+1)} = \theta^{(n)} + B\left.\frac{\partial l}{\partial\theta}\right|_{\theta=\theta(n)}, B = \frac{1}{2}\left(-\frac{\partial^2 l}{\partial\theta\partial\theta^t} + E\left[-\frac{\partial^2 l}{\partial\theta\partial\theta^t}\right]\right)$$

Gilmour 등 (1995)은 세 가지 방법 가운데 AI 방법이 가장 빠른 방법임을 보였고, 따라서 AI 방법을 이용하여 θ를 추정하였다. AI 방법에서 사용하는 AI 함수는 관측정보함수 (observed information function)와 피셔정보함수 (Fisher information function)의 평균이다. 만약 $P = H^{-1} - H^{-1}X(X^t H^{-1} X)^{-1} X^t H^{-1}$이라고 정의하면, REML을 위한 점수 (score) 함수는 다음과 같다:

$$\frac{\partial l_r}{\partial\sigma^2} = -\frac{n-m-1}{2\sigma^2} + \frac{1}{2\sigma^4} y^t P y$$

$$\frac{\partial l_r}{\partial\lambda} = -\frac{1}{2} tr(P\Phi) + \frac{1}{2\sigma^2} y^t P\Phi P y$$

$$\frac{\partial l_r}{\partial\rho_i} = -\frac{1}{2} tr(P\Psi_i) + \frac{1}{2\sigma^2} y^t P\Psi_i P y$$

다음은 관측정보함수이다:

$$-\frac{\partial^2 l_r}{\partial\sigma^4} = -\frac{n-m-1}{2\sigma^4} + \frac{1}{\sigma^6}y^t Py$$

$$-\frac{\partial^2 l_r}{\partial\lambda^2} = -\frac{1}{2}tr(P\Phi P\Phi) + \frac{1}{\sigma^2}y^t P\Phi P\Phi Py$$

$$-\frac{\partial^2 l_r}{\partial\lambda\partial\sigma^2} = \frac{1}{2\sigma^4}y^t P\Phi Py$$

$$-\frac{\partial^2 l_r}{\partial\rho_i\partial\sigma^2} = \frac{1}{2\sigma^4}y^t P\Psi_i Py$$

$$-\frac{\partial^2 l_r}{\partial\rho_i\partial\lambda} = -\frac{1}{2}tr(P\Psi_i P\Phi) + \frac{1}{\sigma^2}y^t P\Psi_i P\Phi Py \quad -\frac{\partial^2 l_r}{\partial\rho_j\partial\rho_i} = -\frac{1}{2}tr(P\Psi_j P\Psi_i) + \frac{1}{\sigma^2}y^t P\Psi_j P\Psi_i Py$$

그리고 피셔정보함수는 아래와 같다:

$$E\left[-\frac{\partial^2 l_r}{\partial\sigma^4}\right] = \frac{n-m-1}{2\sigma^4}$$

$$E\left[-\frac{\partial^2 l_r}{\partial\lambda^2}\right] = \frac{1}{2}tr(P\Phi P\Psi)$$

$$E\left[-\frac{\partial^2 l_r}{\partial\lambda\partial\sigma^2}\right] = \frac{1}{2\sigma^2}tr(P\Phi)$$

$$E\left[-\frac{\partial^2 l_r}{\partial\tau_i\partial\sigma^2}\right] = \frac{1}{2\sigma^2}tr(P\Psi_i)$$

$$E\left[-\frac{\partial^2 l_r}{\partial\rho_i\partial\lambda}\right] = \frac{1}{2}tr(P\Phi P\Psi_i)$$

$$E\left[-\frac{\partial^2 l_r}{\partial\rho_j\partial\tau_i}\right] = \frac{1}{2}tr(P\Psi_i P\Psi_j)$$

따라서 관측정보함수와 피셔정보함수로 평균을 이용하는 AI 방법의 점수함수는 다음과 같다:

$$\mathrm{A}(\sigma^2,\sigma^2)_\mathrm{r} = \frac{1}{2\sigma^6}\mathrm{y^t P y}$$

$$\mathrm{A}(\lambda,\lambda)_\mathrm{r} = \frac{1}{2\sigma^2}\mathrm{y^t P\Phi P\Phi P y}$$

$$\mathrm{A}(\sigma^2,\lambda)_\mathrm{r} = \frac{1}{2\sigma^4}\mathrm{y^t P\Phi P y}$$

$$\mathrm{A}(\sigma^2,\rho_\mathrm{i})_\mathrm{r} = \frac{1}{2\sigma^4}\mathrm{y^t P\Psi_i P y}$$

$$\mathrm{A}(\rho_\mathrm{i},\lambda)_\mathrm{r} = \frac{1}{2\sigma^2}\mathrm{y^t P\Psi_i P\Phi P y}$$

$$\mathrm{A}(\rho_\mathrm{i},\rho_\mathrm{j})_\mathrm{r} = \frac{1}{2\sigma^2}\mathrm{y^t P\Psi_i P\Psi_j P y}$$

위 결과를 이용하여 분산모수를 추정하였다. 전체적인 추정의 순서를 정리하면 Figure 2.1과 같다.

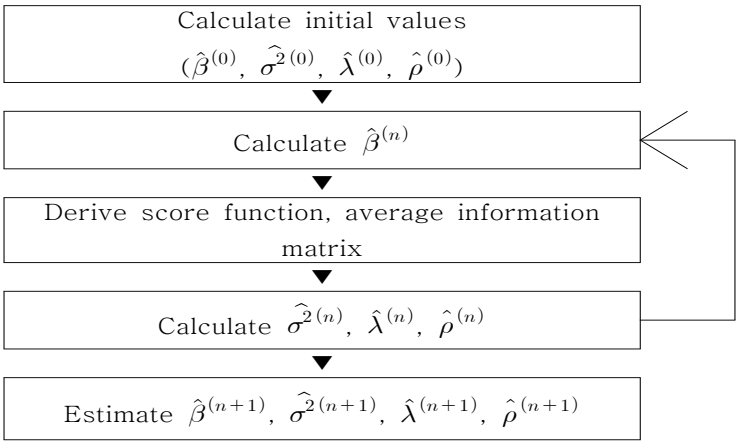

Figure 2.1 Parameter estimation procedure

3. 결과

3.1. 모의실험

부모와 2명의 형제로 구성된 핵가족을 가정하였고, 각 반복실험에서 총 250명의 핵가족 즉 1000명으로 구성된 자료를 생성하여 통계량을 검정하였다. MAF (minor allele frequency)는 0.01로 가정하였고, 유전자형 빈도 (genotype frequency)는 하디-바인베르크 평형 (Hardy-Weinberg equilibrium; HWE) 가정 하에서 계산하였다. 각 핵가족에서 부모의 원인유전자는 다항분포를 가정하여 생성하였고, 자식들의 원인유전자는 멘델의 법칙 (Mendelian transmission)을 가정하여 생성하였다. 표현형은 원인유전자와 β의 곱, 다중유전자 (polygenic) 효과, 환경효과를 합하여 생성하였다. SNP을 제외한 공변량은 존재하지 않는다고 가정하였으며, 환경효과는 $N(0, \sigma^2(I + \rho\Psi))$에서 생성하였다. σ^2은 1.0, ρ는 0, 0.3, 0.5, 그리고

$$\Psi = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

이라고 가정하였다. 만약 h^2는 전체 형질의 분산 대비 다유전자의 효과의 분산의 비율, h_a^2는 전체 형질의 분산 대비 원인유전자 분산의 비율이라고 가정하면, h^2와 h_a^2은 다음과 같이 표현할 수 있다:

$$h^2 = \frac{\sigma_g^2 + 2\beta^2 MAF(1 - MAF)}{\sigma_g^2 + 2\beta^2 MAF(1 - MAF) + \sigma^2}$$

$$h_a^2 = \frac{2\beta^2 MAF(1 - MAF)}{\sigma_g^2 + 2\beta^2 MAF(1 - MAF) + \sigma^2}$$

h^2의 값은 0.3, 0.5, 그리고 h_a^2은 0, 0.005으로 가정을 하였다. 그리고 σ_g^2와 β는 위 식을 통하여 계산하였으며, 다중유전자효과는 $N(0, \sigma_g^2\Phi)$에서 생성하였다.

모의실험 자료를 활용하여 기존 모형 (Model 1)과 새롭게 제안한 모형 (Model 2)를 비교하였다. Model 1과 Model 2의 X로 원인유전자만을 고려하였으며 따라서 X는 $n \times 1$열벡터이다. Model 1은 환경요인으로 인한 가족 간의 공분산을 무시한 모형이다. Model 2는 본 논문에서 제안한 방법으로, 환경요인으로 인한 가족 간의 공분산을 고려하는 모형이다. 두 모형에 관한 자세한 식은 다음과 같다:

$$\text{[Model 1]}$$
$$y = X\beta + b + \epsilon$$
$$b \sim MVN(0, \sigma^2\lambda\Phi), \epsilon \sim MVN(0, \sigma^2 I)$$

$$\text{[Model 2]}$$
$$y = X\beta + b + \epsilon$$
$$b \sim MVN(0, \sigma^2\lambda\Phi), \epsilon \sim MVN(0, \sigma^2(I + \rho\Psi))$$

두 모형을 모의실험자료에 적용하여 결과를 평가하였다. 첫 번째로 새롭게 제안한 Model 2에서 계산된 모수 추정치의 일치성 여부를 평가하였다. 표본크기를 500명부터 3000명까지 증가시키며 추정치가 가정한 값으로 수렴하는지 확인해 보았다. Figure 3.1, 3.2는 표본의 크기별 평균제곱오차 (mean square error; MSE)를 보여주며, y축은 MSE, x축은 표본의 크기를 나타낸다.

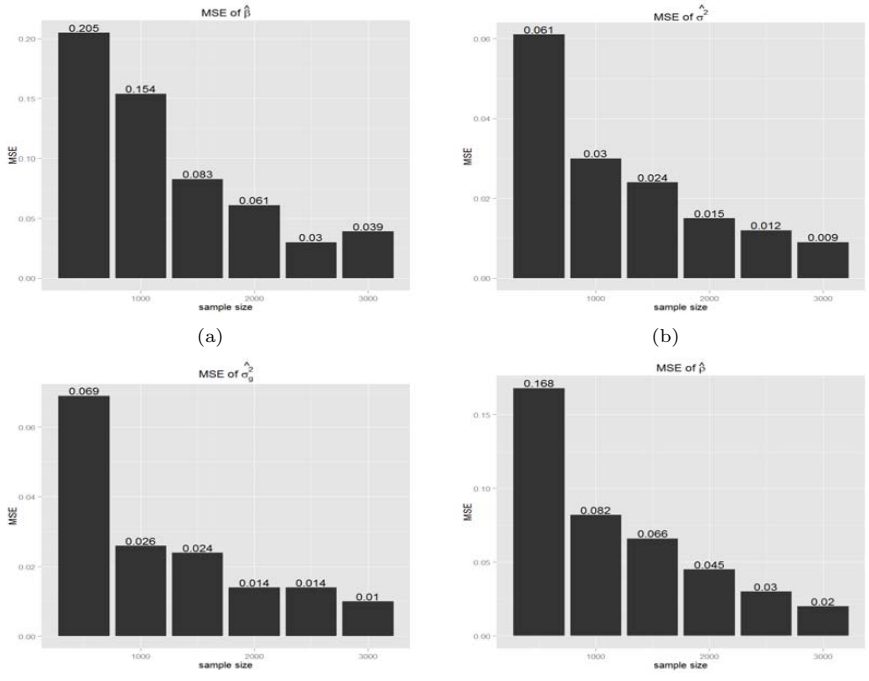

Figure 3.1 Mean square error when β=0

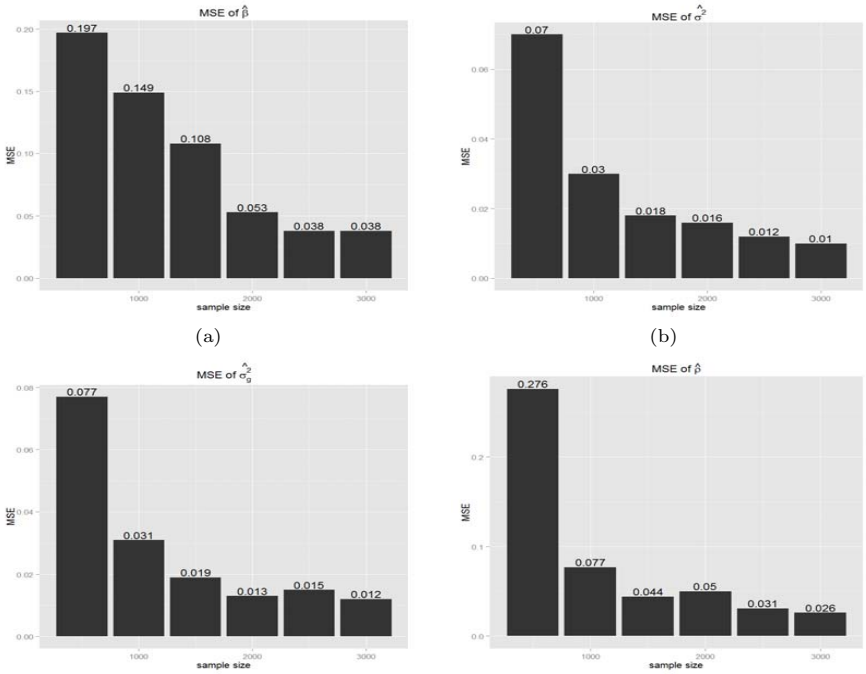

Figure 3.2 Mean square error when β=0.68

Figure 3.1은 $\beta = 0$, $\rho = 0.4$, $h^2 = 0.3$을 가정하였으며, Figure 3.2는 $\beta = 0.68$, $\rho = 0.4$, $h^2 = 0.3$을 가정하였다. Figure 3.1, 3.2는 모두 표본의 크기가 커짐에 따라 MSE값이 작아짐을 보여준다. 따라서 추정된 모수들은 일치성을 만족한다고 볼 수 있다.

다음으로 영가설($\beta = 0$) 하에서 β의 왈드 (Wald) 검정통계량의 경험적 크기 (empirical size; ES)를 구하여, 유의수준 α와 비교하였다. Table 3.1은 $\beta = 0, \rho = 0$일 때의 결과로 ES가 유의수준과 유사함을 보여준다. 따라서 제안한 방법이 통계적으로 타당함을 알 수 있다. Figure 3.3, 3.4는유의확률의 QQ (quantile quantile) 그림으로, Table 3.1의 결과와 마찬가지로 제안한 방법이 통계적으로 타당함을 보여준다. 마찬가지로 $\beta = 0, \rho \neq 0$일 때 ES 결과인 Table 3.2 또한 제안한 방법이 통계적으로 타당함을 보여준다.

Table 3.1 Empirical size for Model 1

Model 1				**Model 2**			
ρ	h^2	α	**ES**	ρ	h^2	α	**ES**
0	0.3	0.01	**0.016**	0	0.3	0.01	**0.015**
		0.05	**0.034**			0.05	**0.054**
		0.1	**0.082**			0.1	**0.096**
		0.2	**0.179**			0.2	**0.185**
	0.5	0.01	**0.013**		0.5	0.01	**0.012**
		0.05	**0.057**			0.05	**0.056**
		0.1	**0.1**			0.1	**0.106**
		0.2	**0.196**			0.2	**0.201**

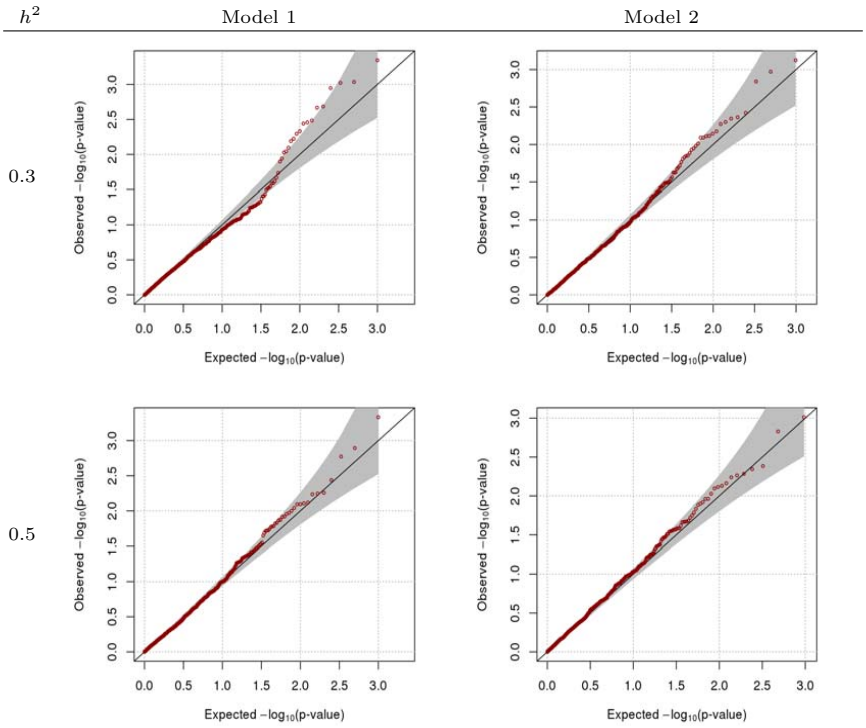

Figure 3.3 QQ-plot when β=0, ρ=0

Table 3.2 Empirical size evaluation for Model 2

Model 1				**Model 2**			
ρ	h^2	α	**ES**	ρ	h^2	α	**ES**
0.4	0.3	0.01	**0.016**	0.4	0.3	0.01	**0.012**
		0.05	**0.034**			0.05	**0.057**
		0.1	**0.082**			0.1	**0.101**
		0.2	**0.179**			0.2	**0.193**
	0.5	0.01	**0.013**		0.5	0.01	**0.013**
		0.05	**0.057**			0.05	**0.06**
		0.1	**0.1**			0.1	**0.119**
		0.2	**0.196**			0.2	**0.205**
0.2	0.3	0.01	**0.016**	0.2	0.3	0.01	**0.015**
		0.05	**0.034**			0.05	**0.06**
		0.1	**0.082**			0.1	**0.099**
		0.2	**0.179**			0.2	**0.191**
	0.5	0.01	**0.013**		0.5	0.01	**0.007**
		0.05	**0.057**			0.05	**0.035**
		0.1	**0.1**			0.1	**0.065**
		0.2	**0.196**			0.2	**0.137**

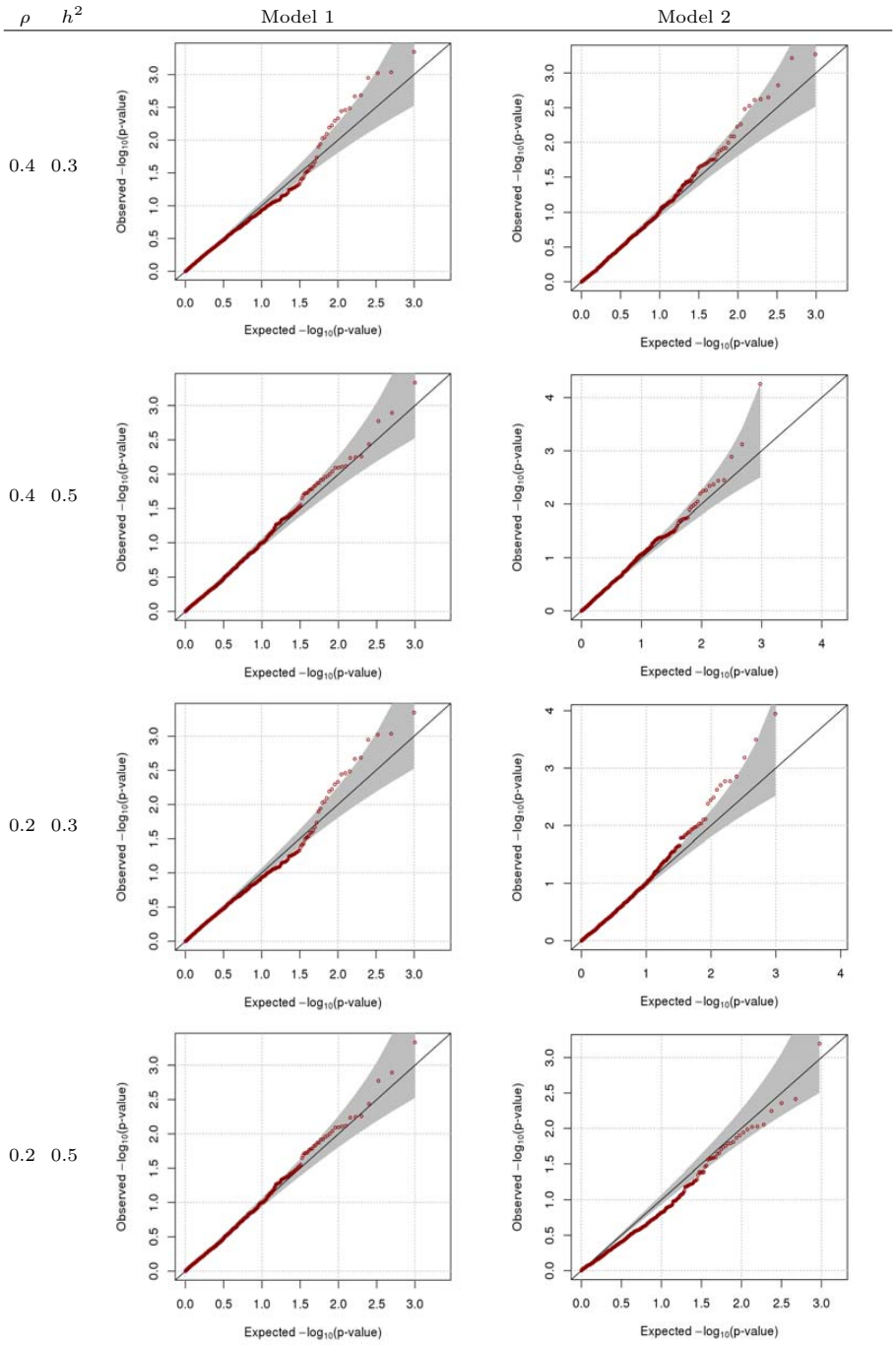

Figure 3.4 QQ-plot when β=0, ρ=0

다음으로 Table 3.3, 3.4는 경험적 검정력 (empirical power; EP)를 보여준다. Table 3.3은 $\beta \neq 0$, $\rho = 0$에서 생성된 모의실험 자료에 대한 검정결과이고, Table 3.4는 $\beta \neq 0$, $\rho \neq 0$일 때의 결과이다. Table 3.3에서는 $\rho = 0$을 가정했기 때문에, Model 1이 정확한 분산공분산 구조를 갖고 있으며, Table 3.4에서는 Model 2가 정확한 분산공분산 구조를 갖고 있다. Table 3.3은 $\rho = 0$일 때, Model 1과 Model 2의 결과가 차이가 없음을 보여주는 반면, Table 3.4는 $\rho \neq 0$일 때는 Model 2의 EP가 Model 1에 비하여 우수함을 보여준다. Table 3.4는 Model 2를 이용할 경우 대략 5% 정도 검정력이 향상됨을 알 수 있다. 따라서 환경요인으로 인한 가족구성원간의 유사성이 기대되는 경우 제안한 모형을 이용함으로써 분석의 효율성을 향상시킬 수 있을 것이다.

Table 3.3 Empirical power when $\rho = 0$

Model 1				**Model 2**			
ρ	h^2	α	**EP**	ρ	h^2	α	**EP**
0	0.3	0.01	**0.08**	**0**	0.3	0.01	**0.08**
		0.05	**0.21**			0.05	**0.22**
		0.1	**0.32**			0.1	**0.34**
		0.2	**0.45**			0.2	**0.45**
	0.5	0.01	**0.05**		0.5	0.01	**0.05**
		0.05	**0.16**			0.05	**0.17**
		0.1	**0.23**			0.1	**0.25**
		0.2	**0.36**			0.2	**0.38**

Table 3.4 Empirical power when $\rho \neq 0$

Model 1				**Model 2**			
ρ	h^2	α	**EP**	ρ	h^2	α	**EP**
0.4	0.3	0.01	**0.294**	0.4	0.3	0.01	**0.336**
		0.05	**0.513**			0.05	**0.566**
		0.1	**0.646**			0.1	**0.663**
		0.2	**0.77**			0.2	**0.759**
	0.5	0.01	**0.285**		0.5	0.01	**0.336**
		0.05	**0.499**			0.05	**0.555**
		0.1	**0.629**			0.1	**0.668**
		0.2	**0.752**			0.2	**0.78**
0.2	0.3	0.01	**0.294**	0.2	0.3	0.01	**0.328**
		0.05	**0.513**			0.05	**0.564**
		0.1	**0.646**			0.1	**0.668**
		0.2	**0.77**			0.2	**0.764**
	0.5	0.01	**0.285**		0.5	0.01	**0.324**
		0.05	**0.499**			0.05	**0.555**
		0.1	**0.629**			0.1	**0.662**
		0.2	**0.752**			0.2	**0.772**

3.2. HTK 코호트 자료 분석

HTK 코호트 자료 기반 전장유전체 분석을 통하여 BMI의 원인유전자를 규명하고자 하였다. HTK 코호트는 530 가족으로 구성되어 있으며, 총 개체수는 2194명이다. 반응변수는 BMI 이고, 환경 변수로 나이와 나이2, 성별을 고려하였다. MAF가 0.05미만이거나 HWE 검정의 유의확률이 10-4보다 작은 SNPs들은 분석에서 제거하였고, 결과적으로 453,376개의 SNPs이 분석에 활용되었다. Figure 3.3은 독립변수 및 반응변수의 히스토그램을 보여준다.

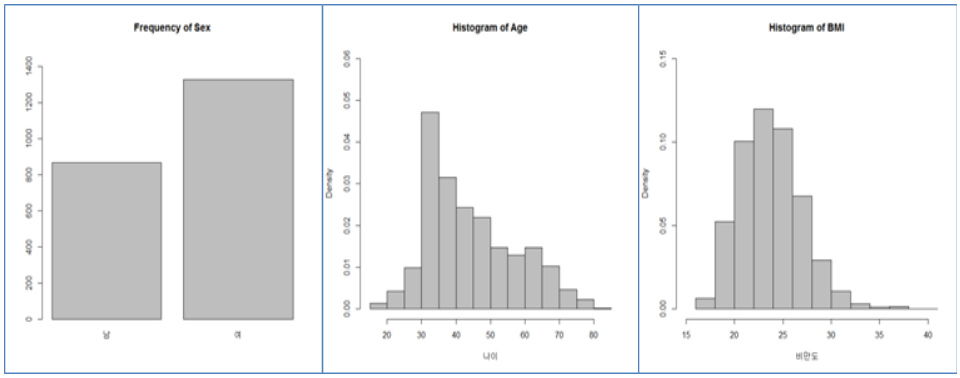

Figure 3.5 Histograms for age, sex, and BMI

Table 3.5는 TWIN 데이터를 본 논문에서 제시하는 Model 1과 Model 2에 적용한 결과를 보여준다. SNPs이 많은 관계로 소프트웨어 EMMAX (Kang 등, 2010)를 이용하여 상위 10개의 SNP을 선별한 뒤에, 제안한 방법으로 분석한 결과를 정리하였다. Table 3.5에서 * 표시는 본페로니 수정을 한 뒤에도 유의한 SNP을 표시한 것이다. BMI는 환경에 의하여 영향을 많이 받는 표현형으로, Model 2는 환경요인으로 인하여 형제간에 유사성이 존재한다고 가정하였다. 분석결과 환경요인으로 인한 형제간의 유사성을 고려한 Model 2에서 SNPs의 p값이 더 작음을 알 수 있었다.

Table 3.5 Association analysis results from TWIN data

SNP	Model 1	Model 2
SNP1	**2.32E-07***	**2.17E-07***
SNP2	1.61E-05	2.50E-05
SNP3	2.33E-05	2.80E-05
SNP4	4.59E-06	2.16E-06
SNP5	5.95E-06	3.87E-06
SNP6	8.08E-06	6.27E-06
SNP7	9.03E-06	7.14E-06
SNP8	**9.26E-07***	**2.09E-07***
SNP9	1.04E-05	5.28E-06
SNP10	9.14E-05	9.83E-05

4. 결 론

현재까지 가족 자료를 이용한 유전연관 분석은 가족구성원 간 표현형의 유사성이 유전적 요인에 의하여 발생한다는 가정 하에 진행되었다. 본 논문에서는 다양한 형태의 분산 공분산 구조를 고려할 수 있는 선형혼합모형과 제안한 모형의 모수를 추정할 수 있는 방법을 제안하였다. 모의실험을 활용하여 영가설이 참일 때 제안한 방법의 검정 통계량이 타당하고, 대립가설이 참일 때 기존의 분석보다 검정력이 우수함을 확인하였다. 또한 실제 데이터에 적용할 때, 좀 더 유의한 결과들을 얻을 수 있음을 BMI의 전장유전체 분석을 통하여 확인하였다. 따라서 제안한 방법을 활용하여 환경적 요인에 의한 가족 구성원간 유사성을 적절히 고려할 때, 원인 유전자를 좀 더 효율적으로 규명할 수 있는 것으로 기대된다.

그러나 제안한 방법은 모수를 추정할 때 기존의 방법보다 많은 계산을 필요로 한다. 예를 들어 샘플의 크기를 n이라 하면, 분산공분산구조의 역행렬 등은 $O(n^3)$으로 증가하고, 따라서 전장유전체 분석의 경우 수 개월의 시간이 필요할 수 있다. 따라서 후속 연구를 통하여 계산량을 줄일 수 있는 방법을 개발할 때 좀 더 효율적인 전장유전체 분석이 가능할 것이다.

References

Corbeil, R. R. and Searle, S. R. (1976). Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics*, **18**, 31-38.

Diggle, P., Heagerty, P., Liang, K. Y. and Zeger, S. (2002). *Analysis of longitudinal data*, 2nd Ed., Oxford University Press, USA.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, **56**, 52-64.

Falconer, D. S., Mackay, T. F. and Frankham, R. (1996). Introduction to quantitative genetics (4th edn). *Trends in Genetics*, **12**, 280.

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.

Jennrich, R. I., and Sampson, P. F. (1976). Newton-Raphson and related algorithms for maximum likelihood variance component estimation. *Technometrics*, **18**, 11-17.

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**, 348-354.

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

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963-974.

Lee, J. (2010). *Genetic variation and diseases*, 2nd Ed., World Science, Korea.

Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, **83**, 1014-1022.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., *et al.* (2009). Finding the missing heritability of complex diseases. *Nature*, **461**, 747-753.

Neudecker, H. and Magnus, J. R. (1999). *Matrix differential calculus with applications in statistics and econometrics*, 2nd Ed., Wiley, New York.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545.

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 B*, **58**, 572.

Stoline, M. R. (1981). The status of multiple comparisons: Simultaneous estimation of all pairwise comparisons in one-way ANOVA designs. *The American Statistician*, **35**, 134-141.

Sung, J., Cho, S. I., Lee, K., Lee, M., Ha, M., Choi, E. Y., Choi, J. S., Kim, H. K., *et al.* (2006), Healthy twin: A twin-family study of Korea–protocols and current status. *Twin Research and Human Genetics*, **9**, 844-848.

Valdar, W., Solberg, L. C., Gauguier, D., Burnett, S., Klenerman, P., Cookson, W. O., Taylor, M. S., Rawlins, J. N. P., Mott, R. and Flint, H. (2006). Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature Genetics*, **38**, 879-887.

Zhou, X. and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, **44**, 821-824.

Journal of the Korean Data & Information Science Society
2014, **25**(5), 1025–1038

http://dx.doi.org/10.7465/jkdi.2014.25.5.1025
한국데이터정보과학회지

Efficient strategy for the genetic analysis of related samples with a linear mixed model[†]

Jeongmin Lim[1] · Joohon Sung[2] · Sungho Won[3]

[1]Chunlab, Inc.
[23]Department of Public Health Science, Seoul National University

Received 30 June 2014, revised 24 July 2014, accepted 5 August 2014

Abstract

Linear mixed model has often been utilized for genetic association analysis with family-based samples. The correlation matrix for family-based samples is constructed with kinship coefficient and assumes that parental phenotypes are independent and the amount of correlations between parent and offspring is same as that of correlations between siblings. However, for instance, there are positive correlations between parental heights, which indicates that the assumption for correlation matrix is often violated. The statistical validity and power are affected by the appropriateness of assumed variance covariance matrix, and in this thesis, we provide the linear mixed model with flexible variance covariance matrix. Our results show that the proposed method is usually more efficient than existing approaches, and its application to genome-wide association study of body mass index illustrates the practical value in real data analysis.

Keywords: Average information method, genome-wide association study, linear mixed model, Newton-Raphson method, restricted maximum likelihood.

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2013R1A1A2010437).
[1] Researcher, Chunlab, Inc., Seoul National University, Seoul 151-742, Korea.
[2] Professor, Department of Public Health Science, Seoul National University, Seoul 151-742, Korea.
[3] Corresponding author: Assistant professor, Department of Public Health Science, Seoul National University, Seoul 151-742, Korea. E-mail: won1@snu.ac.kr