응용통계연구 제 **21**권 *1*호, *2008*년, *pp. 19–33*

다변량 형질의 유전연관성에 대한 주성분을 이용한 회귀방법와 다변량 비모수 추세검정법의 비교

김수영[1] 송혜향[2]

요 약

연속 형질(Quantitative trait)에 영향을 미치는 유전자를 알아내기 위해 형제 쌍의 자료를 수집하여, 주로 이용되는 Haseman과 Elston (1972)의 최소제곱 회귀검정법으로 분석하는데 이는 단일 형질에 대한 분석법이다. 현실적으로 여러 형질들이 복잡하게 단일유전자 좌위(single locus)와 연관되어 있어 함께 수집하게 되는 경우에는, 이러한 연관된 여러 형질을 동시에 분석하는 유전연관성 검정법(linkage test)이 절실히 필요한 실정이다. Amos 등 (1990)은 주성분(principal component) 선형모형을 이용하여 Haseman과 Elston (1972) 방법을 둘 이상의 형질의 다변량 분석법으로 확장시켰다. 그러나 이 검정방법은 통계량의 분포를 알 수 없기에 아직 제 1종 오류가 제대로 통제되지 못하는 문제를 가지고 있다. 본 논문에서는 이러한 다변량 형질 자료의 연관성검정에 있어 단일변량에 대한 비모수 추세검정법을 다변량 자료에 대한 분석법으로 확장시킨 통계량을 사용할 것을 제안한다. Amos 등 (1990)이 제안한 방법과 다변량 추세검정 통계량을 모의실험으로 생성한 연속형 형질자료에 적용하였을 때, 다변량 추세검정 통계량은 Amos 등 (1990) 방법에서의 여러 문제점이 발생되지 않을 뿐만 아니라 모의실험에서 제 1종 오류가 정해진 유의수준에 가까운 것을 확인하였고, 검정력이 더 높음을 볼 수 있었다.

주요용어: Haseman과 Elston (1972)의 회귀분석, Amos 등 (1990)의 회귀분석, 다변량 비모수 추세검정, 유전연관성 검정.

1. 서론

형제 쌍(Sib pair)의 연속 형질(quantitative trait) 자료에 대한 유전연관성 검정(linkage test)에서 일반적으로 쓰이는 Haseman과 Elston (1972)의 최소제곱 회귀분석 방법은 형질에 영향을 미칠 것으로 예상되는 표지유전자 좌위(marker locus)에서 형제 쌍이 공유한 대립유전자(allele) 수가 많을수록 두 형제의 형질이 유사할 것이라는 단순한 원리를 바탕으로 한다. 형제 쌍에서 각 개체의 형질 값을 X_1, X_2라고 한다면, $Y = (X_1 - X_2)^2$을 종속변수로 하고, 두 형제가 부모로부터 물려받은 대립유전자의 공유(identical by descent, IBD)

1) (137-701) 서울시 서초구 반포동 505, 가톨릭대학교 의학통계학과, 석사과정.
E-mail: kimsuyoung@catholic.ac.kr
2) (137-701) 교신저자. 서울시 서초구 반포동 505, 가톨릭대학교 의학통계학과, 교수.
E-mail: hhsong@catholic.ac.kr

정도를 독립변수로 둔 회귀분석에서 음의 기울기 대립가설을 검정함으로써 유전연관성 여부를 밝혀낸다.

다양한 유전적 질환에서 다면발현(pleiotropism)이 관측되고 있는데, 이와 같은 유전적 질환의 발현은 여러 가지 형질로 변환된 후의 결과만으로 파악될 수 있다 (Amos 등, 1990). 이러한 경우에 Amos 등 (1990)은 주요 유전자(major gene)에 의해 동시에 영향을 받는 여러 형질의 정보를 분석에 모두 이용한다면 더욱 검정력이 높아질 것임을 시사하고, 저자들은 Haseman과 Elston (1972)의 회귀분석법을 다변량 형질 자료의 분석에로 확장시켰다. Amos 등 (1990)은 우선 여러 형질 자료와 유전자 IBD와의 선형관계를 최대화하는 주성분 선형 결합변수를 가장 큰 특성근(characteristic root)만을 사용하여 추정하고 Haseman과 Elston (1972)의 회귀식의 기울기를 검정하므로써 유전연관성을 밝힌다. 그러나 Amos 등 (1990)과 Elston 등 (2000)은 이 방법은 제 1종 오류가 정해진 유의수준과 일치하지 않으므로 더욱 연구되어야 한다고 지적하고 있다. 또한 회귀분석 방법은 정규성 가정과 분산의 동질성 가정이 기본이 되지만 두 형제의 형질 수치 차의 제곱은 이러한 가정이 만족되지 않는다. 그러므로 각 개체의 형질의 분포에 대한 어떠한 가정도 요구되지 않는 비모수적 방법이 더욱 적절하다는 것이 언급되었다 (Wan 등, 1997; Wang 등, 1998). 형제 쌍이 공유한 대립유전자 수에 따라 전체 형제 쌍 자료가 세 군으로 분류될 때, 형질 차의 제곱을 순위자료로 변환하여 Kim 등 (2006)은 평균과 분산의 두 가지 측면에서 비모수적 방법, 다시 말하면 Jonckheere (1954)와 Terpstra (1952)의 위치 통계량과 Siegel과 Tukey (1960)의 산포 통계량을 병합한 비모수 추세검정 방법을 제안하였고 정규성 자료라고 간주될 수 없는 상황에서 더욱 검정력이 높음을 밝혔다.

본 논문에서는 형제 쌍의 여러 형질 자료를 동시에 고려하여 유전연관성을 검정하는 비모수적 방법을 제안하고, 또한 이 통계량을 단일형질에 대한 Haseman과 Elston (1972)의 t 검정의 확장인 Amos 등 (1990)의 F 통계량의 검정력과 모의실험으로 비교할 것이다. 다변량 비모수적 추세검정법의 이론적 배경은 방법론에서 설명하게 될 Dietz (1989)의 논문에 기초를 둔다. Dietz (1989)는 Jonckheere (1954)의 감소추세 대립가설 (ordered alternative)의 통계량을 다변량의 경우로 일반화하였다. 계산이 매우 간편한 다변량 추세검정 통계량을 아직까지 형제 쌍의 다형질 유전연관성 검정의 경우에 적용시킨 경우가 없으며, 이에 대한 가능성을 본 논문에서 알아볼 것이다.

본 논문에서 제안한 다변량 추세검정과 Amos 등 (1990)의 주성분 선형모형을 이용한 회귀검정의 효율성을 비교하기 위해서 관심의 대상인 표지유전자 좌위가 형질좌위(trait locus)와 일치한다고 가정하고서 출발한다.

2. 방법

Haseman과 Elston (1972)의 형제 쌍 분석법에 의하면 연속형질 X_i는 다음과 같은 선형모형을 따르는 것을 가정한다.

$$X_i = \mu + g_i + e_i, \qquad i = 1, 2. \tag{2.1}$$

여기서 i는 첫째 또는 둘째 형제를 나타내며, μ는 전체평균이고 g_i와 e_i는 각각 유전적 효과(genetic effect)와 환경적 효과(environment effect)를 나타낸다. 대립유전자 빈도(allele frequency)가 각각 p와 $q(=1-p)$인 두 대립유전자 B, b에 대해, 유전적 효과를 나타내는 g_i는 유전자형이 BB일 경우 a이고, Bb일 경우 d, bb일 경우 $-a$로 정의할 수 있다. Haseman과 Elston (1972) 회귀 분석법은 다음의 모형을 기본으로 한다.

$$E(Y|\pi) = \alpha + \beta\pi. \tag{2.2}$$

Y는 형제 쌍의 연속 형질 수치의 차의 제곱인 $(X_1 - X_2)^2$이고, 독립변수인 π는 형제 쌍이 대립 유전자를 공유한 정도로서 본 논문에서 가정하는 정확한 유전적 정보가 있는 경우에는 π의 가능한 값은 0, 0.5, 1에 한정된다. 만약 가계의 유전적 정보로부터 추정되는 경우에는 π를 $\hat{\pi}$로 대치한다. 한편 편의상 π에 상수 2를 곱하여 0, 1, 2인 대립유전자 공유수로서 제시하기도 하며 서로 상수 관계에 있는 어떤 정의를 선택하여도 검정의 결론에는 변함이 없으므로 본 논문에서도 $\pi = 0, 1, 2$로 정의한다. 이와같이 두 형제가 공유한 대립유전자수를 경우에 따라서는 단순히 IBD로 나타낸다. 유전연관성 검정은 단일형질의 경우에 회귀식 (2.2)에서 음의 기울기에 대한 대립가설을 t 검정으로 알아낸다.

Amos 등 (1990)이 위의 Haseman과 Elston (1972) 회귀식을 다변량의 경우로 확장시킨 자세한 과정은 저자들의 논문으로 돌리며 여기서는 간략한 소개에 그칠 것이다.

2.1. Amos 등 (1990)의 주성분 선형결합 변수를 이용한 회귀분석

Amos 등 (1990)은 Haseman과 Elston (1972)의 단순 회귀의 확장으로서 하나의 유전자 좌위에서 둘 이상의 형질과의 연관성을 동시에 연구하는 방법을 제안했다. 이제 하나의 유전자좌위에 p개의 형질이 연관되었다고 가정하며, i번째 형제의 h번째 형질의 관측값 X_{ih}는 Haseman과 Elston (1972) 방법에서와 유사하게 다음과 같은 선형모형을 가정한다(Amos 등 (1990)의 기호와 일치시키기 위해 p를 그대로 형질의 수로 사용하며, 대립유전자 빈도 p와는 내용상 확실히 구분될 수 있겠다).

$$X_{ih} = \mu_h + g_{ih} + e_{ih}, \qquad i = 1, 2; h = 1, \ldots, p. \tag{2.3}$$

여기서 μ_h는 h번째 형질의 전체평균이고, g_{ih}는 유전적 효과로 유전자형에 따라 다른 값을 가진다. 만약 유전자형이 BB이라면 g_{ih}는 a_h, Bb이라면 d_h, bb이라면 $-a_h$로 정의된다. e_{ih}는 환경적 효과를 나타내며, 각 형제 쌍에 대하여 독립임을 가정한다. Amos 등 (1990)이 제안한 선형모형은 다음과 같다.

$$[c_1(X_{11} - X_{21}) + c_2(X_{12} - X_{22}) + \cdots + c_p(X_{1p} - X_{2p})]^2 = \alpha + \beta\pi + \epsilon. \tag{2.4}$$

결과의 일반화에 영향을 미치지 않는 $E(\epsilon) = 0$의 가정하에서 이 식은 다시 아래와 같이 표현된다.

$$E\left[\sum_{h}^{p} c_h^2(X_{1h} - X_{2h})^2\right] + E\left[\sum_{h=1}^{p-1}\sum_{h'=h+1}^{p} 2c_h c_{h'}(X_{1h} - X_{2h})(X_{1h'} - X_{2h'})\right] = \alpha + \beta\pi. \tag{2.5}$$

각 형제 쌍의 형질 수치 차이로 이루어진 종속변수와 π간의 선형관계를 최대화하는 $p\times 1$ 계수 벡터 $\boldsymbol{c}$의 값을 결정해야 하며 Amos 등 (1990)은 일반적 다변량 회귀분석 방법을 이 경우에 적용시킨다 (Anderson, 1984; Seber, 1988). 유전연관성 검정은 곧 설명하게 되는 식 (2.7)의 R을 최대화하는 주성분 선형함수 $\boldsymbol{d}$를 구하는 것과 같다.

이제 $\boldsymbol{Y}$는 총 n형제 쌍에 대한 형질 차의 제곱과 교차 곱들의 모든 가능한 결합으로 이루어진 종속변수를 나타내며 따라서 $n \times (p^2+p)/2$ 행렬이 되고, $\boldsymbol{\pi}$는 Haseman과 Elston (1972) 방법에서와 같이 형제 쌍이 공유한 IBD로 이루어진 $n \times 1$벡터이다. $\boldsymbol{Y}$와 $\boldsymbol{\pi}$의 공분산 행렬은 다음과 같이 네 부분으로 분할된다.

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{Y},\boldsymbol{\pi}) &= \begin{bmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{21} & \boldsymbol{S}_{22} \end{bmatrix}, \\
\boldsymbol{S}_{11} &= \frac{1}{n-1}[(\boldsymbol{Y}-\mathbf{1}\bar{\boldsymbol{Y}})'(\boldsymbol{Y}-\mathbf{1}\bar{\boldsymbol{Y}})], \\
\boldsymbol{S}_{12} &= \frac{1}{n-1}[(\boldsymbol{Y}-\mathbf{1}\bar{\boldsymbol{Y}})'(\boldsymbol{\pi}-\mathbf{1}\bar{\pi})], \qquad \boldsymbol{S}_{21}=\boldsymbol{S}_{12}', \\
\boldsymbol{S}_{22} &= \frac{1}{n-1}[(\boldsymbol{\pi}-\mathbf{1}\bar{\pi})'(\boldsymbol{\pi}-\mathbf{1}\bar{\pi})].
\end{aligned} \tag{2.6}$$

여기서 $\bar{\boldsymbol{Y}}$는 행렬 $\boldsymbol{Y}$의 n 형제 쌍에 대한 평균인 $1\times(p^2+p)/2$ 벡터이고, $\bar{\pi}$는 IBD의 평균이며, $\mathbf{1}$은 모든 원소가 1인 $n\times 1$ 벡터이다.

유전연관성 검정의 귀무가설은 '$\beta=0$'이고 대립가설은 음의 기울기인 '$\beta<0$'이다. 귀무가설 하에서의 회귀 제곱합 행렬과 오차 제곱합 행렬은 각각 $\boldsymbol{Q}_H = \boldsymbol{S}_{12}{\boldsymbol{S}_{22}}^{-1}\boldsymbol{S}_{21}$, $\boldsymbol{Q}_E = \boldsymbol{S}_{11}-\boldsymbol{Q}_H$이다. 따라서 아래에 제시된 통계량 R을 최대화하는 선형함수 $\boldsymbol{d}$를 구함으로써 유전연관성 검정이 진행되는데 이는 바로 서로 연관되어 있는 변수들을 선형 변환하여 서로 독립인 주성분을 유도하는 것과 같으며 이를 유도할 때 제약조건은 $\boldsymbol{d}'\boldsymbol{d}=1$이다.

$$R = \max_{\boldsymbol{d}} \frac{\boldsymbol{d}'\boldsymbol{Q}_H\boldsymbol{d}}{\boldsymbol{d}'\boldsymbol{Q}_E\boldsymbol{d}}. \tag{2.7}$$

첫째 주성분으로부터 식 (2.5)의 계수 벡터 $\boldsymbol{c}$의 값이 결정된다. 즉, 처음 p개의 원소들은 형질수치 차이 제곱의 계수 ${c_h}^2$이 되며, 나머지 $p(p-1)/2$개의 원소들은 교차 곱의 계수 $2c_hc_{h'}$이다. Amos 등 (1990)은 위의 식 (2.7)의 분자와 분모에 각각의 자유도 p와 $n-p-1$로 나눈 값이 근사적인 F 분포함을 이용하여 검정할 것을 제안하였다. 이 F 분포하는 검정 통계량의 분자와 분모의 자유도는 각각 p와 $n-p-1$이다.

2.2. 비모수 추세검정

2.2.1. 단변량 추세검정 통계량

유전연관성 검정의 핵심은 'IBD가 증가할수록 두 형제가 유사한 형질을 보여 두 형제의 형질 차는 감소한다'는 것이며 이는 바로 감소 추세검정으로 알아볼 수 있다. 비모수 추세검정을 설명하기 위해서 몇가지 기호를 정의한다. IBD가 k인 군의 어떤 형제 쌍에서 첫 번째와 두 번째 형제의 연속 형질 수치를 각각 X_{1k}, X_{2k}라고 할 때 Y_k는 형제 쌍의 형질 차의 제

곱인 $(X_{1k}-X_{2k})^2$이다. Y_k의 분포함수는 각 군의 연속 누적 분포함수가 $F_0(y),\ldots,F_2(y)$일 때, $F_k(y)=F(y-\tau_k)$ $(k=0,1,2)$로 위치 모수에 대한 관계로 표현된다. 유전연관성이 존재하지 않는다면 확률변수 Y_0,Y_1,Y_2는 동일 확률분포하며, 대립가설은 모든 y에 대하여 $F_0(y)\le F_1(y)\le F_2(y)$로 표현되며, 이 관계에서 적어도 하나의 부등식이 존재한다. 그러므로 귀무가설과 대립가설은 다음과 같다.

$$H_0:\tau_0=\tau_1=\tau_2 \qquad vs. \qquad H_1:\tau_0\ge\tau_1\ge\tau_2. \tag{2.8}$$

마찬가지로 이 대립가설의 관계에서도 적어도 하나의 부등식이 존재한다.

이제 단일 형질의 경우 순위에 기초한 비모수 감소추세 검정통계량을 살펴본다. IBD가 증가할수록 형제 쌍의 형질 차의 제곱은 감소추세를 나타내므로 Jonckheere (1954)와 Terpstra (1952)가 제안한 통계량은 다음과 같은 매우 단순한 구성을 가진다.

$$J=\sum_{u=0}^{1}\sum_{v=u+1}^{2}U_{uv}. \tag{2.9}$$

여기서 U_{uv}는 두 군 비교에 대한 순위 자료에 근거한 Mann-Whitney (1947) U 통계량으로서 다음과 같이 정의된다. IBD가 k인 군에 속한 형제 쌍의 수는 n_k이며 이 군에 속한 형제 쌍의 형질 차의 제곱을 구체적으로 Y_{kl} $(k=0,1,2;l=1,\ldots,n_k)$이라 할 때,

$$\begin{aligned} U_{uv} &= \sum_{l=1}^{n_u}\sum_{l'=1}^{n_v}\Phi(Y_{ul},Y_{vl'}), \\ \Phi(a,b) &= \begin{cases} 1, & a<b, \\ \dfrac{1}{2}, & a=b, \\ 0, & a>b \end{cases} \end{aligned} \tag{2.10}$$

이다. 귀무가설 하에서 U_{uv}의 기댓값은 $E(U_{uv})=n_un_v/2$이고, 대표본 근사에서의 분산과 공분산은 Tryon과 Hettmansperger (1973)이 제시한 $\mathrm{Var}(U_{uv})$와 $\mathrm{Cov}(U_{uv},U_{u'v'})$을 이용하여 식 (2.9)의 통계량 J의 평균과 분산을 구하면 다음과 같다. 여기서 N은 $n_0+n_1+n_2$이다.

$$E(J)=\frac{N^2-\sum_{k=0}^{2}n_k^2}{4}, \qquad \mathrm{Var}(J)=\frac{N^2(2N+3)-\sum_{k=0}^{2}n_k^2(2n_k+3)}{72}. \tag{2.11}$$

이제 통계량 $Z=[J-E(J)]/[\mathrm{Var}(J)]^{1/2}$가 대표본 하에서 근사적으로 표준 정규분포함을 이용하여 유전연관성의 감소추세 단측검정이 행해진다.

2.2.2. 다변량 추세검정 통계량

이제 p개의 형질을 함께 사용하여 유전연관성을 알아보는 다변량 감소추세 검정을 설명한다. 본 논문에서 사용하는 다변량 감소추세 검정 통계량은 우선 각 형질에 대해 앞 절에서와 마찬가지로 Jonckheere (1954)와 Terpstra (1952)의 단변량 감소추세 통계량과 분산

을 계산한다. 즉, IBD가 0, 1, 2군의 p개 형질의 자료는 $(\boldsymbol{Y}_{01},\ldots,\boldsymbol{Y}_{0n_0}),(\boldsymbol{Y}_{11},\ldots,\boldsymbol{Y}_{1n_1})$, $(\boldsymbol{Y}_{21},\ldots,\boldsymbol{Y}_{2n_2})$이고, 여기서 $\boldsymbol{Y}_{kl}=(Y_{kl1},\ldots,Y_{klp})'$ $(k=0,1,2;l=1,\ldots,n_k)$은 $p\times 1$ 형질값 벡터이다. $\boldsymbol{Y}_{kl}$의 다변량 누적분포함수는 $F_k(y)$로 표현하고, 각 형질의 주변 분포함수(marginal distribution function)는 $F_k^{(1)}(y),F_k^{(2)}(y),\ldots,F_k^{(p)}(y)$로 표현할 때 p개의 형질에 대한 감소추세 검정의 가설은 모든 y에 대하여 다음과 같다.

$$\begin{aligned} & H_0 : F_0(y)=F_1(y)=F_2(y) \\ vs. \quad & H_1 : F_0^{(h)}(y)\le F_1^{(h)}(y)\le F_2^{(h)}(y), \qquad h=1,\ldots,p. \end{aligned} \tag{2.12}$$

위의 대립가설에서 적어도 하나의 부등식이 성립하며, 각 형질별로 감소추세가 다를 수 있음을 식 (2.12)는 반영하고 있다.

이제 각 형질에 대한 Jonckheere (1954)와 Terpstra (1952)의 감소추세 통계량은 식 (2.9)와 동일하게 다음과 같이 표현된다.

$$J_h=\sum_{u=0}^{1}\sum_{v=u+1}^{2}U_{uv}^{(h)}, \qquad h=1,\ldots,p. \tag{2.13}$$

각 형질에 대한 통계량 J_h의 기댓값과 분산은 위의 식 (2.11)와 동일하다.

이제 여러 형질이 서로 연관되어 있으므로 식 (2.13)에서 계산된 $p\times 1$ 감소추세 통계량 벡터 $(J_1,\ldots,J_p)'$의 공분산을 구하는 문제가 남아 있다. Dietz (1989)는 순위 자료에 근거하여 이 감소추세 통계량 J_h와 $J_{h'}$의 공분산을 제시하고 있는데 이는 Lehmann (1975, p. 370)과 Gibbons (1985, pp. 238–240)에 제시된 식 (2.10)의 $\Phi(Y_{ul},Y_{ul'})$의 공분산을 이용하여 구하였다. 이미 언급하였듯이 여러 형질의 자료를 순위 자료로 변환하는 것이 이 공분산 계산에 우선되는데 이는 곧 설명하게 되는 상관계수 계산에서 요구되는 순위로서, 각 형질별로 따로 따로 모든 군을 통틀어 1부터 $N=\sum_{k=0}^{2}n_k$까지 순위를 매기는 것이다. 이와 같이 생성된 h번째 형질의 m번째 자료의 순위를 R_{mh}로 표현하며 아래 공분산 공식 (2.14)와 (2.15)은 이러한 순위의 함수로 표현되었다.

$$\begin{aligned} \mathrm{Cov}(J_h,J_{h'}) = & \frac{(N+1)\left[\left(N^3-\sum_{k=0}^{2}n_k^3\right)-3\left(N^2-\sum_{k=0}^{2}n_k^2\right)\right]s_{hh'}}{36(N-2)} \\ & +\frac{\left[3N\left(N^2-\sum_{k=0}^{2}n_k^2\right)-2\left(N^3-\sum_{k=0}^{2}n_k^3\right)\right]r_{hh'}}{24(N-2)}. \end{aligned} \tag{2.14}$$

이 공분산 공식에 포함된 $r_{hh'}$와 $s_{hh'}$는 각각 h번째 형질과 h'번째 형질 사이의 켄달(Kendall)과 스피어만(Spearman)의 순위 상관계수로서 다음과 같은 식에 의해 계산된다 (Lehmann 1975, p. 370).

$$r_{hh'} = \frac{2\sum_{m<m'}^{N} \text{sign}[(R_{m'h} - R_{mh})(R_{m'h'} - R_{mh'})]}{N(N-1)},$$
$$s_{hh'} = \frac{3\sum_{m,m',m''}^{N} \text{sign}[(R_{m'h} - R_{mh})(R_{m'h'} - R_{m''h'})]}{N^3 - N}. \quad (2.15)$$

여기서 sign(r)은 아래와 같다.

$$\text{sign}(r) = \begin{cases} 1, & r > 0, \\ 0, & r = 0, \\ -1, & r < 0. \end{cases} \quad (2.16)$$

Dietz (1989)는 앞에서 구한 Jonckheere (1954)와 Terpstra (1952)의 p-변량(p-variate) 통계량과 이의 공분산 행렬로부터 검정 통계량을 제시하였다. 즉, 식 (2.13)에서 $U_{uv}^{(h)}$의 기댓값 $n_u n_v/2$를 뺀 $J_h^c = \sum_{u=0}^{1}\sum_{v=u+1}^{2}[U_{uv}^{(h)} - n_u n_v/2]$ 는 기댓값이 0이 되며 따라서 $p \times 1$벡터 $N^{-3/2}\boldsymbol{J^c} = N^{-3/2}(J_1^c, \ldots, J_p^c)'$은 근사적으로 평균이 $\mathbf{0}$이고 공분산행렬이 $\boldsymbol{V}$인 p-변량 정규분포를 따르게 되며, $\boldsymbol{V}$은 대각원소와 비대각 원소가 각각 $N^{-3}\text{Var}(J_h^c)$, $N^{-3}\text{Cov}(J_h^c, J_{h'}^c)$로 이루어진 행렬이다. 그러므로 p-변량 형질에 대한 다변량 유전연관성 검정방법인 감소추세 비모수 검정통계량은 결과적으로 다음과 같다.

$$D = \frac{N^{-3/2}\mathbf{1}'\boldsymbol{J^c}}{\sqrt{\mathbf{1}'\boldsymbol{V}\mathbf{1}}}. \quad (2.17)$$

여기서 $\mathbf{1}$은 모든 원소가 1인 $p \times 1$벡터이므로 $\mathbf{1}'\boldsymbol{J^c} = \sum_{h=1}^{p} J_h^c$이다. 이 다변량 추세검정 통계량 D는 근사적으로 표준 정규분포를 따른다 (Dietz, 1989, 부록). 대표본의 형제 쌍 다변량 형질자료의 연관성검정은 단측검정으로 행해진다. 본 논문에서 언급한 Amos 등 (1990)이 제안한 회귀분석과 다변량 비모수 추세검정 통계량을 모의실험으로 비교해 볼 것이다.

3. 모의실험

3.1. 모의실험 계획

본 연구는 무작위 교배(random mating) 및 여러 좌위간의 상위성(epistasis)이 없다는 가정과 유전연관성이 균형(linkage equilibrium)이라는 가정 하에서 이루어졌다.

자료 생성과정은 다음과 같다. 먼저, 두 형제의 유전자 IBD가 0, 1, 2가 되는 세 군의 비율이 0.25, 0.5, 0.25가 되도록 랜덤으로 다항분포(trinomial) 자료를 생성한다. 각 개체의 유전자형은 BB, Bb, bb 중 하나로 발생확률은 대립유전자 빈도가 p일 때 각각 p^2, $2pq$, q^2이 된다 (Falconer와 Mackay, 1996). 따라서 두 형제의 가능한 유전자형 조합은 총 9가지이

다. 이제 IBD가 각각 0, 1, 2의 값을 가질 때의 이러한 9가지 가능한 조합의 조건부 확률이 Haseman과 Elston의 1972년도 논문의 표 1에 제시되어 있으며, 위에서 생성된 두 형제의 유전자 IBD값에 따라 이 조건부 확률을 이용하여 다항분포(multinomial)로 형제 쌍의 유전자형 자료를 생성한다. 이 유전자형에 대하여 p개의 형질에 대한 각각의 유전적 효과가 존재하고, 유전자형에서 p번째 형질의 유전적 효과는 각 형질에 대한 유전율(heritability) h^2에 의하여 계산될 수 있으며, 형질이 다음 세대에 유전되는 정도를 나타내는 유전율 h^2는 다음과 같이 정의된다.

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}. \tag{3.1}$$

식 (2.1)의 환경적 효과의 분산인 σ_e^2는 1로 가정하며 유전적 효과 g_i의 분산인 σ_g^2는 $\sigma_a^2+\sigma_d^2$으로서 각각은 다음과 같이 결정된다 (Haseman과 Elston, 1972).

$$\begin{aligned} \sigma_a^2 &= 2pq[a - d(p-q)]^2, \\ \sigma_d^2 &= 4p^2q^2d^2. \end{aligned} \tag{3.2}$$

만약 유전적 모형이 가법모형(additive model)이라면 $d = 0$이고, 우성모형(dominance model)이라면 $d = a$이며, 열성모형(recessive model)이라면 $d = -a$이다. 그러므로 각 형질에 대한 유전율을 h_h^2이라고 한다면, 대립유전자 빈도 p와 h_h^2을 알면 각 형질에 대한 a_h와 d_h가 계산될 수 있다. 다음 단계로 각 개체의 잔차들은 형제간의 상관관계(ρ_s)와 형질간의 상관관계(ρ_t)을 갖도록 생성되어야 한다. 형제간의 상관관계 ρ_s는 과거 연구에 근거하여 0.25로 정하였고, 형질간의 상관계수는 한국인의 본태성 고혈압(essential hypertension)의 다형성(polymorphism) 연구에서 수집된 실제 자료를 토대로 0.5로 가정했다 (Wagener 등, 1982; Bae 등, 2007). 다른 형제 쌍에서의 다른 형질간 상관계수는 0.1로 가정하였다. 임상자료에서는 극단적인 관측치(extreme observations)를 포함하는 정규분포하지 않는 자료가 빈번하기 때문에 비정규분포에 대하여도 모의실험이 실행되어, 각 개체들의 잔차는 정규분포, 로그 정규(log-normal) 분포로 생성하였다. 이렇게 생성된 두 형제의 형질수치의 차를 이용하여 Amos 등 (1990)이 제안한 회귀분석과 다변량 비모수 추세검정 통계량의 검정력을 비교한다. 모의실험의 결과를 제시한 표에서는 전자를 단순히 'Amos 회귀'로, 후자를 '다변량 추세'라 부른다. 또한 다변량 비모수 추세검정에서는 모든 자료의 순위를 이용한 분석방법이므로 비교의 조건을 같게 하기 위해서 원자료에 근거한 모의실험 결과와 순위자료로 변환하여 실시한 모의실험 결과를 모두 제시한다. 가법모형 하에서 모의실험을 하였고, 표본크기 500쌍의 형제에 대해 위에서 설명한 상관계수로 p-변량 형질 자료를 생성하였으며 모의실험의 반복횟수는 1,000번으로 정하였다.

3.2. 모의실험 결과

모의실험은 500쌍의 형제 자료를 생성하여 1,000회 반복하였으나 'Amos 회귀'의 연산과정에서 양정치(positive definite) 행렬이 아닌 경우 계산이 불가능하여 이러한 경우의 자료는 제외하고 1,000회 반복한 결과가 표 3.1부터 표 3.4에 제시되었다. 귀무가설 하에서 결과는 IBD가 0, 1, 2인 세 군에서 형질의 차이가 없음을 의미하는 유전율 $h^2 = 0$인 경우이며

표 3.1: 정규분포 형질 자료에 근거한 두 방법의 제 1종 오류와 검정력

h^2	p	2-변량 형질		3-변량 형질[a]		4-변량 형질[b]	
		다변량 추세	Amos 회귀	다변량 추세	Amos 회귀	다변량 추세	Amos 회귀
0	0.1	0.051	0.060	0.048	0.129	0.044	0.240
	0.3	0.054	0.052	0.046	0.137	0.047	0.246
	0.5	0.052	0.057	0.041	0.126	0.058	0.255
	0.7	0.042	0.050	0.048	0.150	0.047	0.244
	0.9	0.045	0.051	0.062	0.148	0.047	0.256
0.2	0.1	0.447	0.269	0.499	0.222	0.558	0.185
	0.3	0.458	0.235	0.531	0.214	0.653	0.180
	0.5	0.486	0.280	0.535	0.200	0.598	0.186
	0.7	0.477	0.253	0.548	0.218	0.616	0.186
	0.9	0.404	0.225	0.500	0.220	0.550	0.181
0.4	0.1	0.910	0.500	0.936	0.450	0.966	0.430
	0.3	0.971	0.478	0.984	0.507	0.998	0.461
	0.5	0.966	0.487	0.986	0.481	0.992	0.483
	0.7	0.965	0.484	0.986	0.499	0.994	0.473
	0.9	0.896	0.498	0.942	0.472	0.967	0.430
0.6	0.1	0.997	0.493	0.999	0.487	0.999	0.507
	0.3	1.000	0.523	1.000	0.508	1.000	0.474
	0.5	1.000	0.517	1.000	0.494	1.000	0.513
	0.7	1.000	0.499	1.000	0.507	1.000	0.511
	0.9	0.997	0.534	1.000	0.495	0.999	0.501
0.8	0.1	1.000	0.498	1.000	0.491	1.000	0.499
	0.3	1.000	0.498	1.000	0.523	1.000	0.503
	0.5	1.000	0.513	1.000	0.470	1.000	0.509
	0.7	1.000	0.469	1.000	0.521	1.000	0.503
	0.9	0.999	0.491	1.000	0.479	1.000	0.520
0.9	0.1	1.000	0.511	1.000	0.491	1.000	0.499
	0.3	1.000	0.494	1.000	0.503	1.000	0.507
	0.5	1.000	0.506	1.000	0.471	1.000	0.519
	0.7	1.000	0.482	1.000	0.534	1.000	0.485
	0.9	1.000	0.517	1.000	0.513	1.000	0.517

a: Power is computed at the significance level $\alpha = 0.015$.

b: Power is computed at the significance level $\alpha = 0.0025$.

각각의 표에서 첫 부분이 이에 해당한다. 표 3.1부터 표 3.4에 제시된 두 방법론의 제 1종 오류를 살펴 보면 다변량 추세검정에서는 정해진 유의수준 0.05에 상당히 근접한 반면, Amos 회귀는 2-변량 형질의 경우에만 0.05 수준에 가깝고 형질의 수가 증가함에 따라 제 1종 오류가 매우 커진다. 예를 들어서 표 3.1의 정규분포의 경우를 보면 3-변량 형질일 때 제 1종 오류는 0.12-0.15 수준이며 4-변량 형질일 때 제 1종 오류는 0.24-0.26 수준이다. 이는 구해진 검정통계량의 절대값이 매우 작다는 것을 시사한다. 그러므로 두 가지 방법론의 검정

표 3.2: 순위 변환된 정규분포 형질 자료에 근거한 두 방법의 제 1종 오류와 검정력

h^2	p	2-변량 형질		3-변량 형질[a]		4-변량 형질[b]	
		다변량	Amos	다변량	Amos	다변량	Amos
		추세	회귀	추세	회귀	추세	회귀
	0.1	0.045	0.065	0.035	0.127	0.047	0.232
	0.3	0.056	0.048	0.051	0.117	0.044	0.251
0	0.5	0.055	0.057	0.053	0.116	0.063	0.232
	0.7	0.051	0.055	0.040	0.135	0.038	0.234
	0.9	0.057	0.059	0.051	0.147	0.050	0.244
	0.1	0.414	0.234	0.527	0.190	0.563	0.192
	0.3	0.452	0.262	0.554	0.242	0.597	0.205
0.2	0.5	0.505	0.293	0.568	0.250	0.596	0.205
	0.7	0.465	0.233	0.569	0.227	0.594	0.205
	0.9	0.438	0.241	0.497	0.203	0.533	0.160
	0.1	0.909	0.466	0.944	0.425	0.945	0.395
	0.3	0.965	0.475	0.989	0.491	0.989	0.455
0.4	0.5	0.980	0.474	0.982	0.468	0.995	0.490
	0.7	0.974	0.489	0.988	0.500	0.990	0.489
	0.9	0.905	0.462	0.956	0.428	0.953	0.396
	0.1	0.999	0.500	0.999	0.494	1.000	0.486
	0.3	1.000	0.497	1.000	0.505	1.000	0.520
0.6	0.5	1.000	0.492	1.000	0.469	1.000	0.518
	0.7	1.000	0.467	1.000	0.501	1.000	0.481
	0.9	0.995	0.492	1.000	0.499	1.000	0.456
	0.1	1.000	0.494	1.000	0.527	1.000	0.486
	0.3	1.000	0.503	1.000	0.493	1.000	0.484
0.8	0.5	1.000	0.493	1.000	0.506	1.000	0.477
	0.7	1.000	0.493	1.000	0.516	1.000	0.474
	0.9	1.000	0.511	1.000	0.502	1.000	0.486
	0.1	0.999	0.493	1.000	0.484	1.000	0.487
	0.3	1.000	0.488	1.000	0.526	1.000	0.491
0.9	0.5	1.000	0.490	1.000	0.498	1.000	0.487
	0.7	1.000	0.506	1.000	0.496	1.000	0.506
	0.9	1.000	0.492	1.000	0.497	1.000	0.540

a: Power is computed at the significance level $\alpha = 0.015$.

b: Power is computed at the significance level $\alpha = 0.0025$.

력을 동일한 제 1종 오류수준에서 비교하기 위해서 Amos 회귀의 유의수준이 0.05 정도가 되도록 보정해 준 후에 검정력을 비교하였고, 구체적으로는 반복적인 모의실험으로 2-, 3-, 4-변량 형질의 경우에 Amos 회귀의 적절한 유의한계(significance limit)는 각각 0.05, 0.015, 0.0025로 정한 후 검정력을 구하였다.

검정력 결과는 아래의 표 3.1부터 표 3.4의 제 1종 오류 다음에 제시되었다. 검정력을 비교해 보면 각 형제의 형질수치가 정규 분포하는 표 3.1의 경우나 표 3.2의 순위변환된 정규

표 3.3: 로그 정규분포 형질 자료에 근거한 두 방법의 제 1종 오류와 검정력

h^2	p	2-변량 형질		3-변량 형질[a]		4-변량 형질[b]	
		다변량 추세	Amos 회귀	다변량 추세	Amos 회귀	다변량 추세	Amos 회귀
0	0.1	0.042	0.059	0.038	0.119	0.052	0.271
	0.3	0.052	0.049	0.042	0.131	0.057	0.266
	0.5	0.033	0.051	0.041	0.112	0.049	0.251
	0.7	0.055	0.043	0.047	0.092	0.048	0.263
	0.9	0.045	0.041	0.052	0.112	0.048	0.249
0.2	0.1	0.264	0.074	0.285	0.061	0.310	0.065
	0.3	0.308	0.082	0.340	0.073	0.350	0.042
	0.5	0.307	0.082	0.337	0.070	0.335	0.050
	0.7	0.306	0.088	0.306	0.055	0.355	0.050
	0.9	0.271	0.060	0.299	0.057	0.287	0.044
0.4	0.1	0.576	0.061	0.604	0.076	0.652	0.049
	0.3	0.714	0.111	0.753	0.109	0.771	0.075
	0.5	0.718	0.097	0.776	0.096	0.806	0.076
	0.7	0.723	0.085	0.747	0.095	0.807	0.084
	0.9	0.603	0.070	0.643	0.054	0.678	0.041
0.6	0.1	0.758	0.052	0.798	0.044	0.796	0.033
	0.3	0.939	0.110	0.968	0.108	0.972	0.091
	0.5	0.955	0.124	0.970	0.131	0.971	0.108
	0.7	0.948	0.113	0.980	0.128	0.981	0.115
	0.9	0.823	0.068	0.874	0.067	0.882	0.049
0.8	0.1	0.790	0.034	0.797	0.024	0.826	0.024
	0.3	0.995	0.091	0.993	0.123	0.994	0.072
	0.5	0.993	0.148	0.995	0.176	0.995	0.136
	0.7	0.995	0.121	0.998	0.140	0.991	0.155
	0.9	0.906	0.072	0.917	0.070	0.946	0.063
0.9	0.1	0.785	0.010	0.797	0.007	0.806	0.000
	0.3	0.995	0.098	0.996	0.134	1.000	0.084
	0.5	0.999	0.158	0.998	0.163	0.998	0.163
	0.7	0.997	0.121	0.997	0.161	0.998	0.168
	0.9	0.912	0.000	0.922	0.000	0.956	0.000

a: Power is computed at the significance level $\alpha = 0.015$.

b: Power is computed at the significance level $\alpha = 0.0025$.

분포 자료의 경우에 다변량 추세검정이 Amos 회귀에 비하여 검정력이 상당히 높고 가법모형의 가정 하에서 연속형질의 유전율(h^2)이 증가함에 따라 검정력이 증가한다. 또한 다변량 추세검정의 경우에는 다변량 형질의 수가 증가할수록 검정력이 더욱 높게 나타난 반면, Amos 회귀의 경우에는 형질의 수가 증가할수록 검정력이 낮아지는 것으로 나타나 여러 변량 형질의 정보가 보탬이 되지 못하는 모순된 결과를 제시하고 있다. 표 3.3이나 표 3.4에서와 같이 개체의 형질 수치들이 로그 정규분포하는 경우, 또는 순위변환된 로그 정규분포

표 3.4: 순위 변환된 로그 정규분포 형질 자료에 근거한 두 방법의 제 1종 오류와 검정력

h^2	p	2-변량 형질		3-변량 형질[a]		4-변량 형질[b]	
		다변량 추세	Amos 회귀	다변량 추세	Amos 회귀	다변량 추세	Amos 회귀
0	0.1	0.042	0.060	0.056	0.110	0.054	0.232
	0.3	0.055	0.074	0.048	0.124	0.047	0.240
	0.5	0.047	0.053	0.043	0.119	0.046	0.249
	0.7	0.052	0.058	0.044	0.122	0.052	0.252
	0.9	0.061	0.058	0.049	0.125	0.047	0.254
0.2	0.1	0.289	0.246	0.301	0.236	0.313	0.186
	0.3	0.310	0.271	0.342	0.252	0.389	0.224
	0.5	0.296	0.276	0.338	0.265	0.340	0.200
	0.7	0.333	0.278	0.309	0.239	0.366	0.219
	0.9	0.249	0.242	0.300	0.211	0.307	0.185
0.4	0.1	0.595	0.455	0.616	0.449	0.637	0.393
	0.3	0.732	0.496	0.753	0.500	0.771	0.467
	0.5	0.721	0.475	0.779	0.499	0.808	0.495
	0.7	0.716	0.487	0.786	0.509	0.794	0.479
	0.9	0.619	0.434	0.660	0.458	0.693	0.408
0.6	0.1	0.763	0.495	0.788	0.505	0.787	0.479
	0.3	0.929	0.525	0.955	0.486	0.966	0.509
	0.5	0.962	0.519	0.970	0.525	0.973	0.498
	0.7	0.952	0.505	0.977	0.478	0.966	0.512
	0.9	0.810	0.522	0.860	0.491	0.884	0.491
0.8	0.1	0.780	0.471	0.831	0.510	0.814	0.492
	0.3	0.992	0.540	0.990	0.507	0.996	0.497
	0.5	0.997	0.510	0.993	0.524	0.998	0.507
	0.7	0.997	0.503	0.999	0.504	0.999	0.494
	0.9	0.903	0.499	0.926	0.501	0.935	0.481
0.9	0.1	0.795	0.504	0.804	0.524	0.824	0.495
	0.3	0.993	0.505	0.996	0.535	0.999	0.500
	0.5	0.999	0.525	0.998	0.500	0.998	0.494
	0.7	0.994	0.492	0.999	0.510	0.997	0.491
	0.9	0.907	0.492	0.917	0.482	0.948	0.500

a: Power is computed at the significance level $\alpha = 0.015$.

b: Power is computed at the significance level $\alpha = 0.0025$.

자료의 경우에도 마찬가지로 다변량 추세검정이 Amos 회귀에 비하여 검정력이 여전히 높으며, 유전율이 증가함에 따라 검정력 또한 증가하는 경향을 나타낸다. 특히 표 3.3의 경우에 Amos 회귀의 검정력은 매우 낮은 것으로 나타났다.

4. 결론

Amos 등 (1990)은 단일 형질에 대한 Haseman과 Elston (1972)의 단순회귀 방법에서 다변량 형질의 분석 방법을 이끌어 내어 유전적 질환이 둘 이상의 형질에서 발현되는경우에 대한 유전연관성 분석을 가능하게 하였다. 이 방법은 여러 종속변수들의 가능한 모든 결합을 고려하여 IBD와 연관성이 가장 높은 선형 결합을 찾으며 이와 같이 찾은 선형 결합을 종속변수로 두어 다시 Haseman과 Elston (1972)의 회귀분석을 시행한다. 이러한 이유로 Amos 회귀에서 사용한 검정통계량의 정확한 분포를 알 수 없으며 저자들이 추론과정에서 제시한 근사적인 F 분포로의 가정은 아직 더 연구되어져야 한다. 그러므로 Amos 회귀는 다변량 분석이 가능하다는 장점이 있음에도 불구하고 여러 형질 수치로 이루어진 계수 벡터와 유전연관성 선형모형의 기울기를 함께 추정해야 하는 과정에서 실제로 검정통계량의 분포를 쉽게 파악할 수 없으며 따라서 제 1종 오류에서 문제가 있음을 저자들이 시사하면서, F 분포의 옳은 임계값(critical value)이 구해져야 한다고 토의내용에서 언급하고 있다. 반면에 비모수적 다변량 추세검정의 유전연관성 분석으로의 적용은 제 1종 오류가 조정되면서도 검정력이 높다. 특히 극단적인 관측치가 자료에 포함되어 형질 수치가 정규분포하지 않는 경우에 이 비모수적 추세검정 방법이 적절함을 알 수 있다.

상관성이 있는 여러 형질의 공분산을 감안한 비모수적 다변량 추세검정법이 장점이 되는 검정의 상황은 모든 형질이 같은 방향의 대립가설 하에서이며, 다시 말하면 여러 형질의 유의한 정도에서 차이가 있지만 일부 형질에서라도 감소한다는 대립가설이 성립한다면 단변량 추세검정 통계량을 종합한 다변량 감소추세 검정통계량의 검정력이 높게 되며 모의실험 결과가 이를 뒷받침하고 있다. 따라서 다변량 감소추세 검정통계량은 형질의 수가 증가할수록 검정력도 증가하는 것을 모의실험에서 확인할 수 있었다. 이 다변량 감소추세 검정통계량은 한 형질의 경우에 단변량 추세검정 통계량과 일치하는 단순성이 있으며, 이와 달리 Amos 회귀는 단일 형질의 경우에 Haseman과 Elston (1972)의 회귀로 낙착되지 않는다.

참고문헌

Amos, C. I., Elston, R. C., Bonney, G. E., Keats, B. J. B and Berenson, G. S. (1990). A multivariate method for detecting genetic linkage, with application to a pedigree with an adverse lipoprotein phenotype, *The American Journal of Human Genetics*, **47**, 247–254.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd ed., John Wiley & Sons, New York.

Bae, Y., Park, C., Han, J., Hong, Y. J., Song, H. H., Shin, E. S., Lee, J. E., Han, B. G., Jang, Y., Shin, D. J. and Yoon, S. K. (2007). Interaction between GNB3 C825T and ACE I/D polymorphisms in essential hypertension in Koreans, *Journal of Human Hypertension*, **21**, 159–166.

Dietz, E. J. (1989). Multivariate generalizations of Jonckheere's test for ordered alternatives, *Communications in Statistics: Theory and Methods*, **18**, 3763–3783.

Elston, R. C., Buxbaum, S., Jacobs, K. B. and Olson, J. M. (2000). Haseman and Elston revisited, *Genetic Epidemiology,* **19,** 1–17.

Falconer, D. S. and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*, 4th ed., Longman Scientific & Technical, New York.

Gibbons, J. D. (1985). *Nonparametric Statistical Inference*, 2nd ed., Marcel Dekker, New York.

Haseman, J. K. and Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus, *Behavior Genetics*, **2**, 3–19.

Jonckheere, A. R. (1954). A distribution-free k-sample test against ordered alternatives, *Biometrika*, **41**, 133–145.

Kim, M. K., Hong, Y. J. and Song, H. H. (2006). Nonparametric trend statistic incorporating dispersion differences in sib pair linkage for quantitative traits, *Human Heredity*, **62**, 1–11.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*, Holden-Day, San Francisco.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other, *The Annals of Mathematical Statistics*, **18**, 50–60.

Seber, G. A. F. (1988). *Multivariate Observations*, John Wiley & Sons, New York.

Siegel, S. and Tukey, J. W. (1960). A nonparametric sum of ranks procedure for relative spread in unpaired samples, *Journal of the American Statistical Association*, **55**, 429–445.

Terpstra, T. J. (1952). The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking, *Indagationes Mathematicae*, **14**, 327–333.

Tryon, P. V. and Hettmansperger, T. P. (1973). A class of non-parametric tests for homogeneity against ordered alternatives, *The Annals of Statistics*, **1**, 1061–1070.

Wan, Y., Cohen, J. and Guerra, R. (1997). A permutation test for the robust sib-pair linkage method, *Annals of Human Genetics*, **61**, 79–87.

Wang, J., Guerra, R. and Cohen, J. (1998). Statistically robust approaches for sib-pair linkage analysis, *Annals of Human Genetics*, **62**, 349–359.

Wagener, D., Kuller, L., Orchard, T., LaPorte, R., Rabin, B. and Drash, A. (1982). Pittsburgh diabetes mellitus study. II. Secondary attack rates in Families with insulin-dependent diabetes mellitus, *American Journal of Epidemiology*, **115**, 868–878.

[2007년 8월 접수, 2007년 10월 채택]

Comparison of Principal Component Regression and Nonparametric Multivariate Trend Test for Multivariate Linkage

Su-Young Kim[1] Hae-Hiang Song[2]

Abstract

Linear regression method, proposed by Haseman and Elston(1972), for detecting linkage to a quantitative trait of sib pairs is a linkage testing method for a single locus and a single trait. However, multivariate methods for detecting linkage are needed, when information from each of several traits that are affected by the same major gene are available on each individual. Amos *et al.* (1990) extended the regression method of Haseman and Elston(1972) to incorporate observations of two or more traits by estimating the principal component linear function that results in the strongest correlation between the squared pair differences in the trait measurements and identity by descent at a marker locus. But, it is impossible to control the probability of type I errors with this method at present, since the exact distribution of the statistic that they use is yet unknown. In this paper, we propose a multivariate nonparametric trend test for detecting linkage to multiple traits. We compared with a simulation study the efficiencies of multivariate nonparametric trend test with those of the method developed by Amos *et al.* (1990) for quantitative traits data. For multivariate nonparametric trend test, the results of the simulation study reveal that the Type I error rates are close to the predetermined significance levels, and have in general high powers.

Keywords: Haseman and Elston(1972) regression, principal components linear model, multivariate nonparametric trend test, linkage test.

1) Graduate Student, Dept. of Biostatistics, The Catholic University of Korea, Seoul 137-701, Korea. E-mail: kimsuyoung@catholic.ac.kr

2) Corresponding author. Professor, Dept. of Biostatistics, The Catholic University of Korea, Seoul 137-701, Korea.
E-mail: hhsong@catholic.ac.kr