응용통계연구 (2010)
23(5), 845–855
DOI: 10.5351/KJAS.2010.23.5.845

형제 자료에 근거한 유전연관성 추세 검정법의 비교

오영신[1] · 김한상[2] · 송혜향[3]

[1]가톨릭대학교 대학원 의학통계학과, [2]가톨릭대학교 대학원 의학통계학과
[3]가톨릭대학교 대학원 의학통계학과

(2010년 5월 접수, 2010년 8월 채택)

요 약

의학의 여러 분야의 연관성 연구에서 효율성이 높은 방법으로 채택되고 있는 질병-대조 연구계획을 다수 형제 자료에 근거한 유전연관성 연구에 적용하기 위해서는 가족 자료에 근거한 추세 검정통계량이 요구된다. 독립된 개체에 적용하는 Cochran-Armitage 추세 검정통계량은 가족 자료의 경우 제 1종 오류가 보장되지 않으며, 동일 가족의 다수 형제 자료로 인한 공분산을 감안한 추세 검정통계량이 제시되어야 한다. 본 논문에서는 특히 동일 가족의 질병 형제수에 따른 가중을 도입하여 질병과 관련된 마커좌위에서의 유전자형 자료가 수집되는 경우에 검정력이 더욱 높게 되는 검정통계량을 제안한다. 예제 가족 자료로 가중을 고려한 경우와 고려하지 않은 경우의 검정통계량을 계산하여 비교한다.

주요용어: Cochran-Armitage 추세검정, 질병-대조 연구, 유전 연관성, 형제 자료.

1. 서론

의학분야에서 진행되는 질병과 위험인자간의 연관성 연구는 대다수가 질병-대조 연구계획으로 진행되고 있다 (Breslow, 1996). 만약 질병-대조 연구계획에서 위험인자가 순위변수로 측정되어 순위변수값에 따라 질병율의 증가(또는 감소) 추세를 가정할 수 있다면 Cochran-Armitage (Cochran, 1954; Armitage, 1955)의 추세 검정법(trend test)이 일반 연관성 검정법의 검정력보다도 높다. 나날이 급속하게 발전하는 유전학 연구에서도 질병-대조 연구계획은 장점이 있음을 Risch와 Merikangas (1996)는 강조하였다. 추세를 감안한 유전연관성 검정법은 대립유전자(allele) A가 질병과 관련된 대립유전자일 때 유전자형 aa, Aa, AA에 따라, 즉 각 유전자형이 소유한 질병관련 대립유전자 A의 개수 0, 1, 2에 따라 질병에 걸린 비율의 증가 추세를 검정하는 것이다.

Cochran-Armitage (Cochran, 1954; Armitage, 1955) 추세 검정법은 독립된 개체로부터 수집된 자료에 대한 연관성 검정법이지만, 유전연관성 연구는 일반적으로 질병을 가진 개체(proband)를 출발점으로 가족 자료를 수집하게 된다. Risch와 Teng (1998)은 가족 중 다수 형제가 질병을 가진 경우 단 한명의 질병 형제를 가진 가족에 비해 질병을 유발시킨다고 짐작되는 대립유전자 비율이 평균적으로 높아진다고 하였고, Monks 등 (1998)을 비롯한 여러 연구자들은 다수 질병형제가 유전연관성 검정에서 구체적으로 검정력을 증가시킨다고 하였다 (Monks 등, 1998; Risch, 2000; Fingerlin 등, 2004; Li 등, 2006; Kerber 등, 2008). 한편 Gauderman 등 (1999)과 Moore 등 (2005)은 질병 형제자료와는 독립된 대조

[3]교신저자: (137-701) 서울시 서초구 반포동 505, 가톨릭대학교 대학원 의학통계학과, 교수. 인간유전체다형성 연구소. E-mail: hhsong@catholic.ac.kr

군의 자료가 연관성 연구에 더욱 효율적이라고 밝혔다. 이러한 이유로 유전연관성 연구는 다수 질병 형제자료와 더불어 질병 가족자료와는 독립된 대조군 자료를 주로 이용하고 있으며, 본 논문에서 이와같은 질병-대조 자료를 대상으로 한다.

다수의 가족 자료를 포함하는 질병-대조 자료에서 서로 다른 가족 자료는 독립이지만 동일 가족에 속한 형제는 연관되어 있으므로 가족 내 형제간의 공분산을 감안한 추세 검정통계량이 사용되어야 한다. Slager와 Schaid (2001)는 가족 자료에 근거한 추세 검정통계량의 분산 계산에서 가족 관계를 감안한 공분산을 ITO 행렬을 이용하여 구할 것을 제안하였으며, 이 ITO 행렬방법은 Li와 Sacks (1954)가 제안한 것으로 다음 장에서 설명하듯이 친족관계에 있는 두 명 개체의 유전자형의 결합확률분포(joint probability distribution)로부터 유도한 것이다. 질병을 가진 자녀 중심의 가족 자료 수집에서는 다수 질병 자녀의 가족 자료가 일반 모집단에서보다 더욱 많이 수집되므로 질병 자녀수에 따라 가중을 부여하는 Cochran-Armitage 추세 연관성검정 통계량을 개발하게 되면, 이러한 가중 Cochran-Armitage 추세 검정통계량은 단순 Cochran-Armitage 추세 검정통계량보다 더욱 검정력이 높을 것을 예상할 수 있다.

본 논문에서는 가족 자료에서 추출된 질병을 가진 형제 자료에 근거한 가중 Cochran-Armitage 추세 검정통계량을 제안한다. 그러나 Slager와 Schaid (2001)가 분석한 질병-대조 자료에서와 마찬가지로 질병군과 대조군이 서로 독립이지만 질병군에 속한 형제 자료는 연관될 수 있고, 또한 질병을 가지지 않은 대조군에 속한 형제 자료도 연관될 수 있음을 가정한다. 이러한 질병군과 대조군 형제 자료의 확보는 질병을 가진 자녀수 또는 질병을 가지지 않은 자녀수 중 어떤 자녀수가 더욱 많은가에 따라 대조군 또는 질병군에 소속시키는 것으로 가능하다. 즉 질병을 가진 자녀수가 더 많은 가족에서 질병을 가진 자녀의 연관성 연구에서 요구되는 유전자형 자료를 수집하고, 마찬가지로 질병을 가지지 않은 자녀수가 더 많은 가족에서 질병을 가지지 않은 자녀의 유전자형 자료를 수집하는 것이다. 더욱 보편적으로 형제가 아닌 친족의 자료가 형제 자료와 함께 연관성 연구에 포함될 수 있으나 계산상의 복잡성 때문에 본 논문에서는 직접 다루지 않으며 토의에서 언급하게 된다.

논문의 순서로는 독립된 개체 자료에 근거한 Cochran-Armitage 추세 검정통계량을 2장에서 간략히 설명하고 형제간의 공분산을 고려한 추세 검정통계량을 설명한다. 3장에서는 가중 Cochran-Armitage 추세 검정통계량을 독립된 개체 자료에 근거한 경우와 형제간의 공분산을 고려한 경우로 나누어 설명한다. 예제 자료의 분석 결과와 간단한 모의실험 결과를 4장에 제시한다.

2. 단순 Cochran-Armitage 추세 검정

2.1. 독립된 개체 자료의 Cochran-Armitage 추세 검정

Cochran-Armitage (Cochran, 1954; Armitage, 1955) 추세 검정법은 표 2.1과 같이 $n_{0\cdot}$명 대조군과 $n_{1\cdot}$명 질병군의 독립된 각 개체에서 수집한 유전자형에 따라 분류된 도수 n_{ij} $(i = 0, 1;\ j = 0, 1, 2)$에 근거한다. 여기서 $i = 0, 1$은 각각 대조군과 질병군을 나타내며 $j = 0, 1, 2$는 유전자형 aa, Aa, AA를 나타낸다. 독립된 총 개체수는 $n_{\cdot\cdot}$명이며, $n_{0\cdot}$명 대조군과 $n_{1\cdot}$명 질병군은 서로 독립된 다항분포(multinomial distribution)를 따른다.

질병과 관련된 대립유전자가 A일 때 유전자형 aa, Aa, AA에 대해 점수 $x_j = j$ $(j = 0, 1, 2)$를 부여하여 질병율의 추세를 표현한다. Cochran-Armitage 추세 검정통계량 U와, 질병의 유무와 유전자형간에 연관성이 없다는 귀무가설 하에서의 U의 분산은 다음과 같다 (Cochran, 1954; Armitage, 1955).

$$U = \sum_{j=0}^{2} x_j \left(\frac{n_{0\cdot}}{n_{\cdot\cdot}} n_{1j} - \frac{n_{1\cdot}}{n_{\cdot\cdot}} n_{0j} \right), \qquad \hat{V} = \frac{n_{0\cdot} n_{1\cdot} [4n_{\cdot 2}(n_{\cdot\cdot} - n_{\cdot 2} - n_{\cdot 1}) + n_{\cdot 1}(n_{\cdot\cdot} - n_{\cdot 1})]}{n_{\cdot\cdot}^3}, \qquad (2.1)$$

표 **2.1.** 대조군과 질병군에서의 유전자형 도수

	유전자형			합
	aa	Aa	AA	
대조군	n_{00}	n_{01}	n_{02}	$n_{0\cdot}$
질병군	n_{10}	n_{11}	n_{12}	$n_{1\cdot}$
합	$n_{\cdot 0}$	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

여기서 분산 $\hat{V}$은 질병율의 추세에 대한 수치를 구체적으로 $x_j = j$ $(j = 0, 1, 2)$로 둔 경우이다.

이제 연관성 검정은 U의 기대값이 0이므로 대표본 하에서 검정통계량 $U/\sqrt{\hat{V}}$가 근사적으로 정규분포함을 이용하여 단측으로 시행한다. 일부 프로그램 패키지에서는 분산의 분모에서 $n_{\cdot\cdot}^3$ 대신에 $n_{\cdot\cdot}^2(n_{\cdot\cdot} - 1)$을 제시하기도 하는데, 결과적으로는 $n_{\cdot\cdot}^3$을 사용한 경우보다 약간 작은 검정통계량값을 얻게 된다.

Cochran-Armitage 추세 검정통계량에 근거한 연관성 검정은 한편 기울기에 대한 검정으로도 표현될 수 있으며, 다음 장에서 이러한 기울기 표현을 사용하게 된다. 이제 j번째 유전자형에서의 질병율을 π_j라 할 때 유전연관성 검정은 $\pi_j = \alpha + \beta x_j$ $(j = 0, 1, 2)$에서 기울기 $H_0 : \beta = 0$에 대한 검정과 동일하다 (Agresti, 2002, p.181). 기울기 추정량 b와 b의 분산 추정량은 다음과 같다 (Yates, 1948).

$$b = \frac{\sum_{j=0}^{2} x_j n_{1j} - \sum_{j=0}^{2} n_{1j}\overline{x}}{\sum_{j=0}^{2} x_j^2 n_{\cdot j} - n_{\cdot\cdot}\overline{x}^2}, \qquad \mathrm{Var}(b) = \frac{n_{1\cdot} n_{0\cdot}}{n_{\cdot\cdot}^2}\left(\frac{1}{\sum_{j=0}^{2} x_j^2 n_{\cdot j} - n_{\cdot\cdot}\overline{x}^2}\right), \tag{2.2}$$

여기서 $\overline{x} = \sum_{j=0}^{2} x_j n_{\cdot j}/n_{\cdot\cdot}$이다. 앞에서 제시한 검정통계량 U와 기울기 b는 $U = \left[\sum_{j=0}^{2} n_{\cdot j}(x_j - \overline{x})^2\right] b$의 관계로 대응되므로 연관성 검정은 바로 기울기 β가 0인가에 대한 검정과 동일하다. Cochran-Armitage 추세 검정은 근사정규분포하는 $b/\sqrt{\mathrm{Var}(b)}$을 이용하여 단측으로 시행한다.

2.2. 형제간의 공분산을 고려한 Cochran-Armitage 추세 검정

형제간의 공분산을 고려한 Cochran-Armitage 추세 검정통계량은 식 (2.1)의 U 통계량을 사용하며, 분산만이 달라지게 된다. 형제간의 공분산을 구하기 위해 우선 추세를 반영하는 수치 3×1 열(column) 벡터 $\mathbf{x} = (x_0, x_1, x_2)'$를 정의하고, 질병군과 대조군에 속한 각 개체의 유전자형을 각각 $\mathbf{y}_i = (y_{i0}, y_{i1}, y_{i2})'$와 $\mathbf{z}_i = (z_{i0}, z_{i1}, z_{i2})'$의 지시벡터(indicator vector)로 표현하며 만약 질병군에 속한 i번째 개체의 유전자형이 Aa이라면 $\mathbf{y}_i = (0, 1, 0)'$가 된다. 두 형제의 유전자형의 공분산은 $\mathbf{y}_i$와 $\mathbf{y}_j$가 지시변수이므로 $\mathrm{Cov}(\mathbf{y}_i, \mathbf{y}_j) = E(\mathbf{y}_i\mathbf{y}_j') - [E(\mathbf{y}_i)][E(\mathbf{y}_j')]$에서 첫 번째 항은 바로 결합확률분포가 되고, $E(\mathbf{y}_i)$는 유전자형의 주변확률분포(marginal probability distribution)가 된다.

두 친족의 유전자형의 결합확률분포는 한 개체의 유전자형 정보를 알고 있음을 전제로 한 조건부확률을 이용하여 유전자형 결합의 모든 가능한 경우의 확률을 합하여 구한다. 이제 두 형제의 경우를 설명하면, 두 형제는 부모로부터 각각 유전받은 대립유전자를 0, 1, 2개 공유(identical by descent; IBD)할 수 있으며 이 공유수에 따라 조건부확률이 달라진다. 공유수 2, 1, 0개에 대응되는 조건부확률은 각각 다음과 같은 I, T, O 행렬로 제시되며 Li와 Sacks (1954)에 의해 유도되었다. Slager와 Schaid (2001)에서 이 ITO 행렬방법을 이용하여 공분산을 구한다.

$$\begin{matrix}\text{조건 } aa \\ \text{조건 } Aa \\ \text{조건 } AA\end{matrix} \qquad I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad T = \begin{pmatrix} q & p & 0 \\ q/2 & 1/2 & p/2 \\ 0 & q & p \end{pmatrix}, \qquad O = \begin{pmatrix} q^2 & 2pq & p^2 \\ q^2 & 2pq & p^2 \\ q^2 & 2pq & p^2 \end{pmatrix}, \tag{2.3}$$

여기서 p는 일반적으로 채택되는 기호에 따라 질병과 관련되어있다고 생각되는 대립유전자 A의 비율(allele frequency)이며, 따라서 대립유전자 a의 비율은 $q = 1 - p$이다. 예를 들어, 위의 O행렬은 두 형제가 대립유전자를 전혀 공유하지 않은 경우로써 바로 랜덤교배 집단의 두 개체의 경우와 동일하여 조건부확률 행렬은 유전자형 aa, Aa, AA의 확률인 q^2, $2pq$, p^2이 된다. 공유수가 1개 또는 0개에 대응되는 조건부확률도 이와 같은 방식으로 구한다 (Li와 Sacks, 1954 참고). 한편, 두 형제는 랜덤교배하에서는 대립유전자를 0, 1, 2개를 공유하는 확률이 각각 1/4, 1/2, 1/4이 되므로 한 형제의 유전자형 정보를 알고 있을 때 조건부확률은 $I/4 + T/2 + O/4$가 된다. 따라서 두 형제의 유전자형의 결합확률분포는 $P(g_i)(I/4 + T/2 + O/4)$가 되며, 여기서 $P(g_i)$는 다음과 같다.

$$P(g_i) = \begin{pmatrix} p^2 & 0 & 0 \\ 0 & 2pq & 0 \\ 0 & 0 & q^2 \end{pmatrix}. \tag{2.4}$$

이제 $E(\mathbf{y}_i)$는 유전자형의 주변확률분포인 $\mathbf{p} = (p_0, p_1, p_2)'$로 구하며 여기서 p_i $(i = 0, 1, 2)$는 유전자형 aa, Aa, AA 각각의 확률이다. 이를 종합하여, $\mathbf{w}_i$와 $\mathbf{w}_j$가 연관된 두 형제의 유전자형 지시벡터를 나타낼 때 공분산은 다음과 같이 표현된다.

$$\mathrm{Cov}(\mathbf{w}_i, \mathbf{w}_j) = P(g_i)\left(\frac{1}{4}I + \frac{1}{2}T + \frac{1}{4}O\right) - \mathbf{p}\mathbf{p}'. \tag{2.5}$$

따라서 서로 연관된 개체의 공분산을 감안한 Cochran-Armitage 추세 검정통계량 U의 분산은 다음과 같다 (Slager와 Shaid, 2001).

$$\begin{aligned} \mathrm{Var}(U)_{sib} = &\left(0 - \frac{n_{1\cdot}}{n_{\cdot\cdot}}\right)^2 \mathbf{x}'\left[\sum_i \mathrm{Var}(\mathbf{z}_i) + 2\sum_{i<j}\sum \mathrm{Cov}(\mathbf{z}_i, \mathbf{z}_j)\right]\mathbf{x} \\ &+ \left(1 - \frac{n_{1\cdot}}{n_{\cdot\cdot}}\right)^2 \mathbf{x}'\left[\sum_i \mathrm{Var}(\mathbf{y}_i) + 2\sum_{i<j}\sum \mathrm{Cov}(\mathbf{y}_i, \mathbf{y}_j)\right]\mathbf{x}. \end{aligned} \tag{2.6}$$

이 분산공식에서 만약 모든 개체들이 서로 독립인 경우에는 공분산항 $\mathrm{Cov}(\mathbf{y}_i, \mathbf{y}_j)$와 $\mathrm{Cov}(\mathbf{z}_i, \mathbf{z}_j)$이 생략되면서 단순 Cochran-Armitage 추세 검정통계량의 분산공식과 같아진다. 한편 $\mathrm{Var}(\mathbf{y}_i)$와 $\mathrm{Var}(\mathbf{z}_i)$는 다항분포의 분산-공분산행렬로 $\sigma_{ii} = p_i(1 - p_i)$와 $\sigma_{ij} = -p_i p_j$로 구성되며, 여기서 p_i $(i = 0, 1, 2)$는 위에서 정의한 각각 유전자형 aa, Aa, AA의 확률이다. 이제 질병과 관련되어있다고 생각되는 대립유전자 A의 비율인 p와 유전자형 aa, Aa, AA 각각의 확률인 p_i $(i = 0, 1, 2)$의 추정량으로서 $\hat{p} = (n_{\cdot 1} + 2n_{\cdot 2})/(2n_{\cdot\cdot})$과 $\hat{q} = 1 - \hat{p}$ 및 $\hat{p}_0 = \hat{q}^2$, $\hat{p}_1 = 2\hat{p}\hat{q}$, $\hat{p}_2 = \hat{p}^2$을 대입하여 분산추정량을 구한다.

형제간의 공분산을 고려한 단순 Cochran-Armitage 추세 검정통계량은 다음과 같다.

$$Z = \frac{U}{\sqrt{\widehat{\mathrm{Var}}(U)_{sib}}} \sim Z(0, 1). \tag{2.7}$$

3. 가중 Cochran-Armitage 추세 검정

이제 수집된 질병군의 자료가 여러 군으로 분류되어 점진적인 단계를 나타내 보이는 표 3.1과 같은 경우의 추세 검정을 살펴본다. 서로 다른 군의 개체는 서로 독립이다. 유전학 연구에서 다수 질병 자녀의 가족 자료가 수집되어 질병 자녀수에 따라 더욱 세밀한 질병군의 구분이 가능할 때 유전연관성 분석에서 질병 자녀수에 따라 가중을 부여하게 되면 이와 같은 가중 Cochran-Armitage 추세 검정통계량은 단순

표 **3.1.** 대조군과 여러 질병군에 대한 유전자형 도수

	유전자형			합
	aa	Aa	AA	
대조군	n_{00}	n_{01}	n_{02}	$n_{0\cdot}$
질병 I군	n_{10}	n_{11}	n_{12}	$n_{1\cdot}$
질병 II군	n_{20}	n_{21}	n_{22}	$n_{2\cdot}$
질병 III군	n_{30}	n_{31}	n_{32}	$n_{3\cdot}$
합	$n_{\cdot 0}$	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

Cochran-Armitage 추세 검정통계량보다도 더욱 검정력이 높을 것을 예상할 수 있으며, 동일 질병군에 속한 연관된 형제의 공분산을 감안해야 한다. 독립된 개체의 가중 추세 검정통계량을 설명한 후에 형제 자료의 가중 추세 검정통계량을 제시한다.

3.1. 독립된 개체 자료의 가중 Cochran-Armitage 추세 검정

우선 총 $n_{\cdot\cdot}$명의 개체가 서로 독립인 경우의 가중 추세 검정통계량에 대해 알아본다. 대조군을 시작으로 질병의 심각성에 따라 분류된 질병 I, II, III군은 점진적인 단계를 나타내 보인다. 앞에서 설명한 점수 x_j에 추가하여 $y_i = i$ $(i = 0, 1, 2, 3)$로 질병의 심각성에 따른 가중을 표현한다.

양방향 가중 x_j와 y_i를 함께 고려한 가중 Cochran-Armitage 추세 검정통계량은 Yates (1948)의 방법에서 참고하였고, 또한 Cochran (1954)도 이를 언급하였다. 양방향 가중을 고려한 기울기 b와 b의 분산 추정량은 다음과 같다 (Yates, 1948).

$$b_w = \frac{\sum_{j=0}^{2} x_j \left(\sum_{i=0}^{3} y_i n_{ij} \right) - \left(\sum_{i=0}^{3} y_i n_{i\cdot} \right) \overline{x}}{\sum_{j=0}^{2} x_j^2 n_{\cdot j} - n_{\cdot\cdot} \overline{x}^2}, \qquad \mathrm{Var}(b_w) = \frac{\sum_{i=0}^{3} y_i^2 n_{i\cdot} - n_{\cdot\cdot} \overline{y}^2}{n_{\cdot\cdot} \left(\sum_{j=0}^{2} x_j^2 n_{\cdot j} - n_{\cdot\cdot} \overline{x}^2 \right)}, \tag{3.1}$$

여기서 $\overline{x} = \sum_{j=0}^{2} x_j n_{\cdot j} / n_{\cdot\cdot}$이고, $\overline{y} = \sum_{i=0}^{3} y_i n_{i\cdot} / n_{\cdot\cdot}$이다. 이제 가중 Cochran-Armitage 추세 검정은 $b_w / \sqrt{\mathrm{Var}(b_w)}$이 대표본하에서 근사적으로 정규분포함을 이용하여 단측으로 시행한다. 위의 식 (3.1)에 $y_0 = 0$, $y_1 = y_2 = y_3 = 1$의 가중값을 대입하게 되면 식 (2.2)에 제시된 대조군과 질병군의 두 군의 경우의 기울기 b와 b의 분산 추정량과 동일하게 된다.

위의 식 (3.1)에 제시된 기울기로 표현된 검정통계량은 다음과 같은 가중 Cochran-Armitage 추세 검정통계량으로도 표현됨을 유추할 수 있다.

$$U_w = \sum_{j=0}^{2} x_j \left(\sum_{i=0}^{3} y_i n_{ij} \right) - \left(\sum_{i=0}^{3} y_i n_{i\cdot} \right) \overline{x}, \qquad \hat{V_w} = \frac{\left(\sum_{j=0}^{2} x_j^2 n_{\cdot j} - n_{\cdot\cdot} \overline{x}^2 \right) \left(\sum_{i=0}^{3} y_i^2 n_{i\cdot} - n_{\cdot\cdot} \overline{y}^2 \right)}{n_{\cdot\cdot}}. \tag{3.2}$$

3.2. 형제간의 공분산을 고려한 가중 Cochran-Armitage 추세 검정

가중 Cochran-Armitage 추세 검정통계량은 우선 동일 형제수 가족 자료를 대상으로 하며 본 논문에서는 세 명의 자녀가 있는 가족 자료로 국한하여 설명한다. 이러한 세 자녀 가족의 자료에 근거하여 표 3.1에 제시된 질병 I, II, III군은 질병에 걸린 형제수가 1명, 2명, 3명인 경우가 된다. 따라서 대조군을 시작으로 여러 질병군에 걸쳐 점진적인 단계를 나타내 보인다. 유전자형 aa, Aa, AA에 점수

$x_j = j\,(j = 0, 1, 2)$을 부여하여 질병율의 추세를 표현하며, $y_i = i\,(i = 0, 1, 2, 3)$로 질병에 걸린 형제수에 따른 가중을 표현한다.

연관된 개체 자료에서의 분산 역시 단순 Cochran-Armitage 추세 검정과 마찬가지로 모든 개체들이 서로 독립인 경우의 분산에 연관된 형제들의 공분산항을 추가하여 분산을 계산한다. 서로 연관된 개체의 공분산을 감안한 가중 Cochran-Armitage 추세 검정의 U_w의 분산은 다음과 같다.

$$\begin{aligned}\mathrm{Var}(U_w)_{sib} = \; & (0-\overline{y})^2\,\mathbf{x}'\left[\sum_i \mathrm{Var}(\mathbf{z}_i) + 2\sum_{i<j}\sum \mathrm{Cov}(\mathbf{z}_i, \mathbf{z}_j)\right]\mathbf{x} \\ & + \sum_{k=1}^{3}(k-\overline{y})^2\,\mathbf{x}'\left[\sum_i \mathrm{Var}(\mathbf{y}_{ki}) + 2\sum_{i<j}\sum \mathrm{Cov}(\mathbf{y}_{ki}, \mathbf{y}_{kj})\right]\mathbf{x}. \qquad (3.3)\end{aligned}$$

기호는 2.2절의 것과 동일하며, 단지 k명 질병형제군에 속한 개체의 유전자형을 $\mathbf{y}_{ki} = (y_{ki0}, y_{ki1}, y_{ki2})'$ 지시벡터로 구분하여 표시하였다. 두 형제의 유전자형의 공분산은 2.2절에서와 마찬가지로 Li와 Sacks (1954)의 ITO 행렬방법을 이용하여 유전자형의 결합확률분포로부터 계산한다. 식 (3.3)의 유도과정을 부록에 제시하였다.

형제간의 공분산을 고려한 가중 Cochran-Armitage 추세 검정통계량은 다음과 같다.

$$Z_w = \frac{U_w}{\sqrt{\widehat{\mathrm{Var}}(U_w)_{sib}}} \sim Z(0,1). \qquad (3.4)$$

4. 예제 자료

위에서 제시한 추세 검정통계량과 가중 추세 검정통계량을 계산하기 위해서 임의로 세 명의 자녀를 가진 자료를 생성하여 분석하였다. 대조군에 30가족 59명, 질병군에 30가족 51명, 합해서 총 60가족 110명의 유전자형 자료가 분석에 사용되었다. 대조군과 질병군은 서로 독립이며, 그러나 질병군에는 동일 가족의 형제 자료가 포함되어 있고 대조군의 경우도 마찬가지이다. 유전자형에 따른 대조군과 질병군 가족의 형제 자료가 표 4.1과 4.2에 제시되었다. 표 4.1에 제시된 계수는 식 (3.3)에서 연관된 형제로 인해 추가되는 공분산항의 개수이다.

모든 검정통계량 값은 SAS 9.1 프로그램을 이용하여 계산하였다. 식 (2.1)에 제시된 U 검정통계량의 값은 7.04이며 식 (3.2)에 제시된 U_w 검정통계량의 값은 19.55이다. 또한 표 4.2의 자료로부터 유전자형 aa, Aa, AA 각각의 확률인 $p_i\,(i = 0, 1, 2)$의 추정량은 $\hat{p}_0 = 64/110 = 0.5818$, $\hat{p}_1 = 36/110 = 0.3273$, $\hat{p}_2 = 10/110 = 0.0909$이다. 질병과 관련되어 있다고 생각되는 대립유전자 A의 비율 추정량은 $\hat{p} = 56/220 = 0.2545$이며, 이러한 추정량들을 이용하여 식 (2.5)는 다음과 같이 구해진다.

$$\widehat{\mathrm{Cov}}(\mathbf{w}_i, \mathbf{w}_j) = \begin{pmatrix} 0.0847 & -0.0670 & -0.0439 \\ -0.0670 & 0.1187 & 0.0005 \\ -0.0439 & 0.0005 & 0.0172 \end{pmatrix}. \qquad (4.1)$$

이제 공분산 행렬을 이용하여 식 (2.6)에 적용하여 계산된 분산 추정값은 $\mathrm{Var}(U)_{sib} = 17.6945$이고, 식 (3.3)에 적용하여 계산된 분산 추정값은 $\mathrm{Var}(U_w)_{sib} = 108.9924$이다. 따라서 연관된 형제들의 공분산을 고려한 분석결과는 식 (2.7)에 의한 단순 Cochran-Armitage 추세 검정통계량 Z값이 $1.67\,(p = 0.05)$이며, 식 (3.4)에 의한 가중 Cochran-Armitage 추세 검정통계량의 Z값은 $1.87\,(p = 0.03)$으로 가중의 경우가 더욱 유의하다.

표 **4.1.** 대조군과 질병군의 가족 내 형제의 유전자형 자료

가족	대조군 유전자형			형제수	계수	가족	질병군 유전자형			형제수	계수
	aa	Aa	AA				aa	Aa	AA		
1	1	1	0	2	1	31	0	2	0	2	1
2	2	1	0	3	3	32	0	1	0	1	0
3	0	2	0	2	1	33	0	0	1	1	0
4	1	1	0	2	1	34	1	2	0	3	3
5	1	1	1	3	3	35	1	1	0	2	1
6	2	1	0	3	3	36	0	1	0	1	0
7	2	0	0	2	1	37	1	0	0	1	0
8	1	0	0	1	0	38	1	0	1	2	1
9	2	0	0	2	1	39	1	0	0	1	0
10	1	0	0	1	0	40	0	1	1	2	1
11	1	1	0	2	1	41	1	0	0	1	0
12	1	0	0	1	0	42	0	1	0	1	0
13	0	1	0	1	0	43	1	0	0	1	0
14	2	0	0	2	1	44	1	0	2	3	3
15	1	2	0	3	3	45	1	1	0	2	1
16	1	1	0	2	1	46	1	0	0	1	0
17	1	0	0	1	0	47	1	1	1	3	3
18	2	0	0	2	1	48	2	1	0	3	3
19	1	1	1	3	3	49	1	0	0	1	0
20	1	1	0	2	1	50	0	1	0	1	0
21	2	1	0	3	3	51	1	0	0	1	0
22	0	1	0	1	0	52	1	0	0	1	0
23	2	0	0	2	1	53	1	1	1	3	3
24	1	0	0	1	0	54	2	0	0	2	1
25	2	0	0	2	1	55	1	1	0	2	1
26	2	0	1	3	3	56	1	0	0	1	0
27	2	0	0	2	1	57	0	1	0	1	0
28	2	0	0	2	1	58	2	1	0	3	3
29	2	0	0	2	1	59	0	2	0	2	1
30	0	1	0	1	0	60	2	0	0	2	1
합	39	17	3	59	36	합	25	19	7	51	27

표 **4.2.** 대조군과 질병군에서의 유전자형 도수

	유전자형			합
	aa	Aa	AA	
대조군	39	17	3	59
질병군	25	19	7	51
1명 형제군	9	5	1	15
2명 형제군	8	8	2	18
3명 형제군	8	6	4	18
합	64	36	10	110

간단한 모의실험으로 붓스트랩(Bootstrap) p값을 계산하여 단순 Cochran-Armitage 추세 검정통계량과 가중 Cochran-Armitage 추세 검정통계량을 비교해 본다. 붓스트랩 절차는 다음과 같다. 표 4.1에

표 **4.3.** 붓스트랩 p값 결과

시행	단순 C-A	가중 C-A
1	0.038	0.018
2	0.043	0.031
3	0.041	0.020
4	0.051	0.031
5	0.037	0.019
6	0.038	0.019

제시된 자료를 토대로 각 가족에 대해 형제수를 고정한 상태에서 나올 수 있는 가능한 모든 유전자형 분포를 생성한 후, 이로부터 각 가족을 대표하는 한 분포를 랜덤추출하며 이를 모든 가족에 대해 시행한다. 랜덤 추출을 1000회 반복하며, 매 회마다 생성된 각 자료에 대해 단순 Cochran-Armitage(CA) 추세 검정통계량과 가중 Cochran-Armitage(CA) 추세 검정통계량을 계산한 후, 표 4.1 자료의 검정통계량 값(단순 CA 추세 검정통계량 $Z = 1.67$, 가중 CA 추세 검정통계량 $Z = 1.87$)보다 큰 값을 가진 경우 수를 1000회로 나누어 붓스트랩 p값을 구한다. 이러한 과정을 총 6회 반복한 결과가 표 4.3에 제시되었다. 표 4.3의 붓스트랩 p값을 살펴보면 모든 시행에서 가중 Cochran-Armitage 추세 검정통계량의 p값이 단순 검정통계량의 경우보다 더욱 유의함을 알 수 있다.

5. 토의

본 논문에서는 질병을 가진 자녀 중심의 가족 자료 수집에서 다수 질병 자녀수에 따라 유전연관성의 정도에 대해 가중을 부여하는 Cochran-Armitage 추세 연관성 검정통계량을 제안하였고, 이러한 검정통계량에는 동일 가족 자료에서 추출된 형제 자료의 연관성을 고려한 분산 추정량이 필수적이다. 예제 자료에 적용하여 검정법을 비교한 결과, 연관된 형제의 공분산을 감안한 가중 검정통계량에 의한 분석결과가 이러한 형제의 공분산을 고려하지 않은 경우보다도 더욱 유의하였다. 가족 자료의 연관성을 감안한 가중 Cochran-Armitage 추세 검정법은 질병군과 대조군 형제 자료뿐만 아니라 더욱 보편적으로 형제가 아닌 다른 여러 친족의 자료가 포함되어도 분석할 수 있으며 단지 두 개체의 유전연관성에 따라 식 (2.5)의 결합확률분포인 첫번째 항에서 ITO 행렬과 곱해진 계수 (1/4, 1/2, 1/4)가 친족의 관계에 따라 변해야 한다 (Li와 Sacks, 1954). 즉, 충분한 수의 대가족 자료를 수집할 수 있다면 대가족 내의 관계에 따라 공분산을 구할 수가 있고, 대가족 내의 질병에 걸린 개체의 수에 따라 여러가지로 가중을 주는 방법이 가능할 수 있으며 장래의 연구 과제가 될 수 있다.

부록

식 (3.3)의 유도과정은 다음과 같다.

$$\begin{aligned} U_w &= \sum_{j=0}^{2} x_j \left(\sum_{i=0}^{3} y_i n_{ij} \right) - \left(\sum_{i=0}^{3} y_i n_{i\cdot} \right) \overline{x} \\ &= \sum_{j=0}^{2} x_j \left(n_{1j} + 2n_{2j} + 3n_{3j} \right) - \left(n_{1\cdot} + 2n_{2\cdot} + 3n_{3\cdot} \right) \overline{x} \\ &= \left(\sum_{j=0}^{2} x_j n_{1j} - n_{1\cdot} \overline{x} \right) + 2 \left(\sum_{j=0}^{2} x_j n_{2j} - n_{2\cdot} \overline{x} \right) + 3 \left(\sum_{j=0}^{2} x_j n_{3j} - n_{3\cdot} \overline{x} \right) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=0}^{2} x_j n_{1j} - \frac{n_{1\cdot}}{n_{\cdot\cdot}} \sum_{j=0}^{2} x_j n_{0j} - \frac{n_{1\cdot}}{n_{\cdot\cdot}} \sum_{j=0}^{2} x_j n_{1j} - \frac{n_{1\cdot}}{n_{\cdot\cdot}} \sum_{j=0}^{2} x_j n_{2j} - \frac{n_{1\cdot}}{n_{\cdot\cdot}} \sum_{j=0}^{2} x_j n_{3j} \\
&\quad + 2 \left(\sum_{j=0}^{2} x_j n_{2j} - \frac{n_{2\cdot}}{n_{\cdot\cdot}} \sum_{j=0}^{2} x_j n_{0j} - \frac{n_{2\cdot}}{n_{\cdot\cdot}} \sum_{j=0}^{2} x_j n_{1j} - \frac{n_{2\cdot}}{n_{\cdot\cdot}} \sum_{j=0}^{2} x_j n_{2j} - \frac{n_{2\cdot}}{n_{\cdot\cdot}} \sum_{j=0}^{2} x_j n_{3j} \right) \\
&\quad + 3 \left(\sum_{j=0}^{2} x_j n_{3j} - \frac{n_{3\cdot}}{n_{\cdot\cdot}} \sum_{j=0}^{2} x_j n_{0j} - \frac{n_{3\cdot}}{n_{\cdot\cdot}} \sum_{j=0}^{2} x_j n_{1j} - \frac{n_{3\cdot}}{n_{\cdot\cdot}} \sum_{j=0}^{2} x_j n_{2j} - \frac{n_{3\cdot}}{n_{\cdot\cdot}} \sum_{j=0}^{2} x_j n_{3j} \right) \\
&= \left(1 - \frac{n_{1\cdot}}{n_{\cdot\cdot}} \right) \sum_{j=0}^{2} x_j n_{1j} - \frac{n_{1\cdot}}{n_{\cdot\cdot}} \sum_{j=0}^{2} x_j n_{0j} - \frac{n_{1\cdot}}{n_{\cdot\cdot}} \sum_{j=0}^{2} x_j n_{2j} - \frac{n_{1\cdot}}{n_{\cdot\cdot}} \sum_{j=0}^{2} x_j n_{3j} \\
&\quad + 2 \left(1 - \frac{n_{2\cdot}}{n_{\cdot\cdot}} \right) \sum_{j=0}^{2} x_j n_{2j} - \frac{2n_{2\cdot}}{n_{\cdot\cdot}} \sum_{j=0}^{2} x_j n_{0j} - \frac{2n_{2\cdot}}{n_{\cdot\cdot}} \sum_{j=0}^{2} x_j n_{1j} - \frac{2n_{2\cdot}}{n_{\cdot\cdot}} \sum_{j=0}^{2} x_j n_{3j} \\
&\quad + 3 \left(1 - \frac{n_{3\cdot}}{n_{\cdot\cdot}} \right) \sum_{j=0}^{2} x_j n_{3j} - \frac{3n_{3\cdot}}{n_{\cdot\cdot}} \sum_{j=0}^{2} x_j n_{0j} - \frac{3n_{3\cdot}}{n_{\cdot\cdot}} \sum_{j=0}^{2} x_j n_{1j} - \frac{3n_{3\cdot}}{n_{\cdot\cdot}} \sum_{j=0}^{2} x_j n_{2j} \\
&= \left[0 - \left(\frac{n_{1\cdot} + 2n_{2\cdot} + 3n_{3\cdot}}{n_{\cdot\cdot}} \right) \right] \sum_{j=0}^{2} x_j n_{0j} + \left[1 - \left(\frac{n_{1\cdot} + 2n_{2\cdot} + 3n_{3\cdot}}{n_{\cdot\cdot}} \right) \right] \sum_{j=0}^{2} x_j n_{1j} \\
&\quad + \left[2 - \left(\frac{n_{1\cdot} + 2n_{2\cdot} + 3n_{3\cdot}}{n_{\cdot\cdot}} \right) \right] \sum_{j=0}^{2} x_j n_{2j} + \left[3 - \left(\frac{n_{1\cdot} + 2n_{2\cdot} + 3n_{3\cdot}}{n_{\cdot\cdot}} \right) \right] \sum_{j=0}^{2} x_j n_{3j} \\
&= \mathbf{x}' \left[(0 - \overline{y}) \sum_i \mathbf{z}_i + (1 - \overline{y}) \sum_i \mathbf{y}_{1i} + (2 - \overline{y}) \sum_i \mathbf{y}_{2i} + (3 - \overline{y}) \sum_i \mathbf{y}_{3i} \right]
\end{aligned}
$$

따라서,

$$
\begin{aligned}
\mathrm{Var}(U_w)_{sib} =& \mathrm{Var} \left(\mathbf{x}' \left[(0 - \overline{y}) \sum_i \mathbf{z}_i + (1 - \overline{y}) \sum_i \mathbf{y}_{1i} + (2 - \overline{y}) \sum_i \mathbf{y}_{2i} + (3 - \overline{y}) \sum_i \mathbf{y}_{3i} \right] \right) \\
=& (0 - \overline{y})^2 \mathbf{x}' \mathrm{Var} \left(\sum_i \mathbf{z}_i \right) \mathbf{x} + (1 - \overline{y})^2 \mathbf{x}' \mathrm{Var} \left(\sum_i \mathbf{y}_{1i} \right) \mathbf{x} \\
&+ (2 - \overline{y})^2 \mathbf{x}' \mathrm{Var} \left(\sum_i \mathbf{y}_{2i} \right) \mathbf{x} + (3 - \overline{y})^2 \mathbf{x}' \mathrm{Var} \left(\sum_i \mathbf{y}_{3i} \right) \mathbf{x} \\
=& (0 - \overline{y})^2 \mathbf{x}' \left[\sum_i \mathrm{Var}(\mathbf{z}_i) + 2 \sum_{i<j} \sum \mathrm{Cov}(\mathbf{z}_i, \mathbf{z}_j) \right] \mathbf{x} \\
&+ (1 - \overline{y})^2 \mathbf{x}' \left[\sum_i \mathrm{Var}(\mathbf{y}_{1i}) + 2 \sum_{i<j} \sum \mathrm{Cov}(\mathbf{y}_{1i}, \mathbf{y}_{1j}) \right] \mathbf{x} \\
&+ (2 - \overline{y})^2 \mathbf{x}' \left[\sum_i \mathrm{Var}(\mathbf{y}_{2i}) + 2 \sum_{i<j} \sum \mathrm{Cov}(\mathbf{y}_{2i}, \mathbf{y}_{2j}) \right] \mathbf{x} \\
&+ (3 - \overline{y})^2 \mathbf{x}' \left[\sum_i \mathrm{Var}(\mathbf{y}_{3i}) + 2 \sum_{i<j} \sum \mathrm{Cov}(\mathbf{y}_{3i}, \mathbf{y}_{3j}) \right] \mathbf{x}
\end{aligned}
$$

$$= (0-\overline{y})^2 \mathbf{x}' \left[\sum_i \mathrm{Var}(\mathbf{z}_i) + 2 \sum_{i<j} \sum \mathrm{Cov}(\mathbf{z}_i, \mathbf{z}_j) \right] \mathbf{x}$$
$$+ \sum_{k=1}^{3} (k-\overline{y})^2 \mathbf{x}' \left[\sum_i \mathrm{Var}(\mathbf{y}_{ki}) + 2 \sum_{i<j} \sum \mathrm{Cov}(\mathbf{y}_{ki}, \mathbf{y}_{kj}) \right] \mathbf{x}.$$

참고문헌

Armitage, P. (1955). Tests for linear trends in proportions and frequencies, *Biometrics*, **11**, 375–386.

Breslow, N. E. (1996). Statistics in epidemiology: The case-control study, *Journal of the American Statistical Association*, **91**, 14–28.

Cochran, W. G. (1954). Some methods for strengthening the common chi-squared test, *Biometrics*, **10**, 417–451.

Fingerlin, T. E., Boehnke, M. and Abecasis, G. R. (2004). Increasing the power and efficiency of disease-marker case-control association studies through use of allele-sharing information, *American Journal of Human Genetics*, **74**, 432–443.

Gauderman, W. J., Witte, J. S. and Thomas, D. C. (1999). Family-based association studies, *Journal of the National Cancer Institute Monographs*, **26**, 31–37.

Kerber, R. A., Amos, C. I., Yeap, B. Y., Finkelstein, D. M. and Thomas, D. C. (2008). Design considerations in a sib-pair study of linkage for susceptibility loci in cancer, *BMC Medical Genetics*, **9**, 64.

Li, C. C. and Sacks, L. (1954). The derivation of joint distribution and correlation between relatives by the use of stochastic matrices, *Biometrics*, **10**, 347–360.

Li, M., Boehnke, M. and Abecasis, G. R. (2006). Efficient study designs for test of genetic association using sibship data and unrelated cases and controls, *American Journal of Human Genetics*, **78**, 778–792.

Monks, S. A., Kaplan, N. L. and Weir, B. S. (1998). A comparative study of sibship tests of linkage and/or association, *American Journal of Human Genetics*, **63**, 1507–1516.

Moore, R. M., Pinel, T., Zhao, J. H., March, R. and Jawaid, A. (2005). Selecting cases from nuclear families for case-control association analysis, *BMC Genetics*, **6(Suppl I)**, S105.

Risch, N. (2000). Searching for genetic determinants in the new millennium, *Nature*, **405**, 847–856.

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

Risch, N. and Teng, J. (1998). The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. I. DNA pooling, *Genome Research*, **8**, 1273–1288.

Slager, S. L. and Schaid, D. J. (2001). Evaluation of candidate genes in case-control studies: A statistical method to account for related subjects, *American Journal of Human Genetics*, **68**, 1457–1462.

Yates, F. (1948). The analysis of contingency tables with groupings based on quantitative characters, *Biometrika*, **35**, 176–181.

Comparison of Trend Tests for Genetic Association with Sibship Data

Young-Sin Oh[1] · Han-Sang Kim[2] · Hae-Hiang Song[3]

[1]Department of Biostatistics, Graduate School, The Catholic University of Korea
[2]Department of Biostatistics, Graduate School, The Catholic University of Korea
[3]Department of Biostatistics, Graduate School, The Catholic University of Korea

(Received May 2010; accepted August 2010)

Abstract

Extensively used case-control designs in medical studies can also be powerful and efficient for family association studies as long as an analysis method is developed for the evaluation of association between candidate genes and disease. Traditional Cochran-Armitage trend test is devised for independent subjects data, and to apply this trend test to the biologically related siblings one has to take into account the covariance among related family members in order to maintain the correct type I error rate. We propose a more powerful trend test by introducing weights that reflect the number of affected siblings in families for the evaluation of the association of genetic markers related to the disease. An application of our method to a sample family data, in addition to a small-scale simulation, is presented to compare the weighted and unweighted trend tests.

Keywords: Cochran-Armitage trend test, case-control designs, genetic association, sibship data.

[3]Corresponding author: Professor, Department of Biostatistics, Graduate School, Integrated Research Center for Genome Polymorphism, The Catholic University of Korea, 505 Banpo-Dong, Seocho-Gu, Seoul 137-701, Korea. E-mail: hhsong@catholic.ac.kr