응용통계연구 제 **21**권 *1*호, *2008*년, *pp. 1–17*

Kruglyak과 Lander의 유전연관성 비모수 방법과 반복 자료를 고려한 가중 회귀분석법의 비교

최은경[1] 송혜향[2]

요 약

형제 쌍(sibpair)의 연속형 형질(continuous traits) 자료를 이용한 유전연관성 검정법(linkage test)으로서 Haseman과 Elston (1972)의 최소제곱(ordinary least square, OLS) 회귀분석법이 주로 사용된다. 비모수적 방법으로서 제시된 Kruglyak과 Lander (1995)의 검정통계량은 Haseman과 Elston (1972)의 방법에 대응되는 방법처럼 보이지만 실제로는 매우 다르다. 본 논문에서는 Kruglyak와 Lander (1995)의 검정통계량과 Haseman과 Elston (1972)의 검정통계량의 관계를 설명하고 모의실험으로 두 검정통계량의 검정력을 비교한다. 유전연관성에 사용되는 형제 자료의 특징은 한정된 설명변수의 값에 매우 많은 자료가 반복(replicated)되었다는 점이며, 이러한 반복 자료에 더욱 적절한 가중 회귀분석법을 제안한다. 가중 회귀분석법의 효율성을 정규분포 또는 정규분포가 아닌 연속형 형질 모의실험 자료로 알아본 결과 형제 쌍 자료의 유전연관성 검정에서 가중 회귀분석법이 다른 검정법들보다도 검정력이 높음을 확인하였다.

주요용어: Haseman과 Elston의 회귀분석법, Kruglyak와 Lander의 비모수 통계량, 형제 쌍 자료, 유전연관성 검정.

1. 서론

형제 쌍(sibpair) 자료를 이용한 유전연관성 검정(linkage test) 방법으로서 Haseman과 Elston (1972)의 최소제곱 회귀분석법이 가장 빈번하게 사용되고 있는데, 이는 늦은 연령에서 발병하는 질병(late-onset disease)에 대한 연구에서 환자의 부모가 생존치 않는 경우에 형제 쌍의 자료를 이용하여 손쉽게 유전연관성을 알아 볼 수 있다는 장점 때문이기도 하다. Haseman과 Elston (1972) 회귀검정의 원리는 부모로부터 유전 받은 형제들의 대립 유전자 공유(identical by descent, IBD)의 정도가 증가하면 할수록 두 형제의 형질은 비슷하여 형질 차(trait difference)의 제곱이 작아진다는 것이다. Kruglyak와 Lander (1995)는 이에 대응하는 비모수적 방법을 제안하였는데 저자들은 Haseman과 Elston (1972)의 최소제곱 회귀분석법과의 관계를 충분히 설명치 않았다. 본 논문에서 두 방법의 관계를 알아보고자 한다.

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

Haseman과 Elston (1972) 회귀분석의 특징은 첫째로 형제 쌍이 공유한 대립 유전자수에 대한 정확한 정보가 있는 경우에 유전자 공유수(동일하게 IBD 수라 부르겠다)가 0, 1, 2의 세 수준에 불과하므로 회귀분석에서 설명변수가 되는 IBD 수에 종속변수인 형제 쌍의 형질 차의 제곱이 과도 반복(replicated)된 점이다. 더욱이 세 수준에 반복수가 균등하지 않고 전체 표본수가 크게 되면 IBD 수 0, 1, 2에 자료가 근사적으로 1:2:1의 비율로 반복되게 나타난다. 둘째로 회귀분석에서 요구되는 기본 가정인 각 설명변수의 수준에서 종속변수의 분산이 동일한 σ^2라는 성질이 성립되지 않는다는 점이다 (Wang 등, 1998). 더욱이 정규분포하는 형질의 형제 쌍 자료에서 차의 제곱의 분산 σ^2는 유전연관성이 존재한다면 IBD 수가 0, 1, 2로 증가할수록 감소 경향을 보인다 (Kim 등, 2006). 다시 말해서 여러 측면에서 분산은 일정한 값 σ^2을 가진다는 가정에서 어긋난다. 모형 설정이 잘못된 경우에는 일반적으로 s^2은 σ^2을 과대추정(overestimate)하게 되어 회귀분석의 결론에 있어서 오류의 가능성이 높아진다 (강근석과 김충락, 2001). 이런 이유로 이를 보완한 방법이나 비모수적 방법이 적절함을 알 수 있다.

Kruglyak와 Lander (1995)의 비모수적 방법은 우선 형제 쌍으로부터 구한 형질 차의 제곱의 자료를 순위(rank)로 바꾼다. 이제 IBD 수 0, 1, 2의 세 군의 위치 모수에 대한 검정이 바로 Kruskal과 Wallis (1952)의 비모수 분산분석인데 Kruglyak와 Lander (1995)는 세 군의 순위합을 -1, 0, 1의 가중으로 직선적인 추세에서 가장 차이가 크게 드러나는 IBD 수 0군과 2군의 순위합 비교에 대한 통계량을 제시하였다. 특이하게도 전체 자료에 근거한 회귀분석과 비교하여 대략 절반의 자료에 해당하는 IBD 수가 0과 2인 군만의 자료에 근거한 검정 통계량의 결과가 간혹 더욱 유의한 경우가 있는데, 그러한 이유 중 하나는 여러 군 순위합 비교 후에 다중비교로서 IBD 수 0군과 2군의 대비(contrast)를 검정하는 절차를 거치지 않고 Kruglyak와 Lander (1995)는 곧바로 0군과 2군의 대비만의 검정통계량을 제시하였기 때문이다. 따라서 Haseman과 Elston (1972) 통계량과 Kruglyak와 Lander (1995) 통계량의 검정력 우열을 쉽게 가릴 수 없으며, 우리는 본 논문에서 이 두 검정법을 자세히 비교할 것이다. 이러한 과정에서 Haseman과 Elston (1972)의 통계량에서 요구되는 동일 분산의 가정에 대한 문제점을 보완한 가중 회귀분석법이 유전연관성 검정에서 어떻게 유용할 것이며 이 방법의 적용으로 검정력이 과연 증가할 것인지를 알아본다. 모의실험으로 Haseman과 Elston (1972) 회귀분석과 Kruglyak와 Lander (1995)의 비모수적 방법 및 가중 회귀분석법의 검정력을 비교해 볼 것이다.

2. 방법

Haseman과 Elston (1972)의 형제 쌍 분석법은 질병에 영향을 미친다고 짐작되는 유전자 좌위(marker locus)를 확인하기 위해 연속 형질 자료를 수집하여 이러한 연속 형질과 유전자 좌위 사이의 연관성을 회귀 모형으로 알아보는 분석법이다. 이를 우리는 단순히 Haseman과 Elston (1972)의 회귀검정법이라 부르겠다. Haseman과 Elston (1972)의 회귀모형은 다음과 같다.

$$E(Y_j|x_j) = \alpha + \beta x_j, \qquad x_j = 0, 1, 2;\ j = 1, \ldots, n. \tag{2.1}$$

여기서 x_j는 j번째 형제 쌍의 유전자 공유수로서, $x_j = 0$이면 형제 쌍은 서로 공유하는 유전자가 없고, $x_j = 2$이면 형제 쌍은 동일 유전자형(genotype)을 가진다. 또한 j번째 형제 쌍에서 첫 번째 형제와 두 번째 형제의 연속 형질 값을 각각 Y_{1j}와 Y_{2j}로 정의할 때 Y_j는 두 형제의 형질 수치의 차 제곱(square of the difference)으로 다음과 같다.

$$Y_j = (Y_{1j} - Y_{2j})^2. \tag{2.2}$$

즉, 회귀분석의 기본 자료는 서로 독립인 (x_j, Y_j), $(j = 1, \ldots, n)$이다. Kruglyak와 Lander (1995)의 비모수적 분석법을 설명하고 Haseman과 Elston (1972) 회귀검정법과 어떻게 연관되는지 알아보겠다.

2.1. Kruglyak와 Lander (1995)의 비모수적 방법

Kruglyak와 Lander (1995)의 유전연관성을 알아보는 비모수적 방법은 서로 독립인 전체 형제 자료 $Y_j (j = 1, \ldots, n)$의 순위 $r(j)$를 구하는 것으로 출발한다 (Kruglyak와 Lander(1995)는 두 형제의 연속 형질의 절대값 차(absolute difference)의 순위로 정의하지만 제곱의 순위와 같다). Kruglyak와 Lander (1995)는 이러한 순위를 두 형제의 유전자 공유 정도의 함수 $f(2\pi_j)$로 가중시켜 합한 통계량 $X_{KL} = \sum_{j=1}^{n} r(j) f(2\pi_j)$을 제안하였다. 여기서 π_j는 형제 쌍이 대립 유전자를 공유한 정도로 정의하며 공유한 IBD 수의 정확한 정보가 있는 경우에 π_j의 가능한 값은 0, 0.5, 1이 된다. 그러므로 $2\pi_j$는 형제 쌍이 공유하는 IBD 수가 되어 0, 1, 2의 값이 된다. 가법모형(additive model) 하에서 Kruglyak와 Lander (1995)는 가중 $f(2\pi_j)$를 구체적으로 $f(2) = -1, f(1) = 0$ 그리고 $f(0) = 1$로 정하고 단측 검정을 시행한다(본 논문에서는 Haseman과 Elston (1972)의 회귀분석에서 음의 기울기에 대한 단측 검정을 시행하는 것과 같은 방향을 유지하기 위해 $f(2) = 1, f(1) = 0, f(0) = -1$로 정한다). 즉, Kruglyak와 Lander (1995)의 통계량은 $X_{KL} = R_2 - R_0$이다. 여기서 R_i는 IBD 수가 i인 군의 순위합이다. 유전연관성이 없다는 귀무가설 하에서 통계량 X_{KL}의 기대값은 0이고, Kruglyak와 Lander (1995)는 분산으로서 $V = n(n+1)(2n+1)/12$을 제안하였다. 따라서 $Z_{KL} = X_{KL}/\sqrt{V}$이 근사적으로 표준 정규분포를 따름을 이용하여 유전연관성을 검정하게 된다. 이 분산에 대해서 곧 살펴볼 것이다.

2.2. Haseman-Elston (1972)과 Kruglyak-Lander (1995) 검정 통계량의 관계

두 방법을 비교하기 위해서는 원 자료(raw data)에 근거하는 Haseman과 Elston (1972) 방법과 순위 자료(rank data)에 근거하는 Kruglyak와 Lander (1995)의 통계량의 차이를 잠시 접어 두고서, 두 방법 모두 원 자료에 적용하는 경우로 비교를 시작하겠다.

Dietz (1989)는 회귀 모형 식 (2.1)에서 기울기 β의 최소제곱 추정량 $\hat{\beta}_{LS}$를 단지 두 개의 자료 수치로 구성된 최소 부분군(subset)의 기울기의 병합으로 표현하고 있는데, 이를 유전연관성 검정에 이용하게 되면 $\hat{\beta}_{LS}$는 매우 단순한 형태로 표현될 수가 있다. 우선 자료 (x_j, Y_j), $(j = 1, \ldots, n)$에 근거하여 다음과 같은 부분군의 기울기 $\hat{S}_{ij}$를 정의하며 이러한 부분군의 수는 N개이다. 만약 x_i의 값들이 모두 다르다면 $N = \binom{n}{2}$이 된다.

$$\hat{S}_{ij} = (Y_j - Y_i)/(x_j - x_i), \qquad i < j,\ x_i \neq x_j. \tag{2.3}$$

이제 최소제곱 추정량 $\hat{\beta}_{LS}$는 최소 부분군 기울기 $\hat{S}_{ij}$들의 가중평균으로 구해진다 (Dietz , 1989).

$$\hat{\beta}_{LS} = \frac{\sum_{i<j} \omega_{ij} \hat{S}_{ij}}{\sum_{i<j} \omega_{ij}}. \tag{2.4}$$

x_i가 서로 동일 값을 가지는 최소 부분군의 경우에는 ω_{ij}가 0이 되어 기울기를 계산하지 않게 된다.

2.2.1. 유전 연관성 검정의 기울기 추정량

유전연관성을 알아보는 Haseman과 Elston (1972) 회귀검정법은 일반적인 회귀분석법의 예외적인 경우로서 독립변수가 가능한 값은 세 개뿐이며 구체적으로 IBD 수 0, 1, 2이다. 따라서 각 IBD 수 0, 1, 2에 종속변수 Y_j가 수없이 반복된 경우이다. 본 논문에서는 독립변수의 각 값을 x_0, x_1, x_2로 또는 IBD 수 0, 1, 2로 표현한다. 각 독립변수의 값에 종속변수의 반복수를 $n_0, n_1, n_2 (n = n_0 + n_1 + n_2)$라 정한다. 전체 자료가 서로 독립인 (x_j, Y_j), $(j = 1, \ldots, n)$로 구성될 때 위의 식 (2.4)에서 x_i와 x_j가 함께 동일하여 ω_{ij}가 서로 같은 값을 가지는 최소 부분군들을 종합하여 정리하게 되면 크게 세 덩어리의 부분군으로 다시 표현될 수 있다.

우선 설명에 필요한 몇 가지 기호를 정의한다. IBD 수가 0, 1, 2(즉 독립변수가 x_0, x_1, x_2)인 자료를 (x_0, Y_i), $(i = 1, \ldots, n_0)$; (x_1, Y_{n_0+j}), $(j = 1, \ldots, n_1)$; $(x_2, Y_{n_0+n_1+k})$, $(k = 1, \ldots, n_2)$로 표현하면, 이들의 연속 형질 값의 평균은 $\bar{Y}_0 = 1/n_0 \sum_{i=1}^{n_0} y_i$, $\bar{Y}_1 = 1/n_1 \sum_{j=1}^{n_1} y_{n_0+j}$, $\bar{Y}_2 = 1/n_2 \sum_{k=1}^{n_2} y_{n_0+n_1+k}$이다. 이제 큰 세 덩어리의 부분군을 정의하여 IBD 수가 0과 1인 값만을 가지는 자료로 구성된 '01' 부분군, IBD 수가 1과 2인 값만을 가지는 자료로 구성된 '12' 부분군, 그리고 IBD 수가 0과 2인 값만을 가지는 자료로 구성된 '02' 부분군이다. x_i와 x_j인 값을 가지는 자료만으로 이루어진 부분군의 최소제곱 기울기 추정량은 다음과 같이 표현된다.

$$\begin{aligned} \hat{\beta}_{ij} &= \frac{[n_i n_j (x_j - x_i)(\bar{Y}_j - \bar{Y}_i)]/(n_i + n_j)}{n_i n_j (x_j - x_i)^2/(n_i + n_j)}, \\ &= \frac{\bar{Y}_j - \bar{Y}_i}{x_j - x_i}, \qquad i < j. \end{aligned} \tag{2.5}$$

그러므로 각 부분군의 기울기 $\hat{\beta}_{01}, \hat{\beta}_{12}, \hat{\beta}_{02}$는 부분군에 포함된 IBD 수가 서로 다른 두 군의 Y_j의 평균 차로 단순하게 표현되었다. 이제 '01', '12', '02' 부분군의 가중 $\omega_{01}, \omega_{12}, \omega_{02}$를 구해본다. 서로 다른 x값을 가지는 자료만이 부분군 기울기 계산에 포함되므로 $\omega_{01} = n_0 n_1 (x_1 - x_0)^2$, $\omega_{12} = n_1 n_2 (x_2 - x_1)^2$, $\omega_{02} = n_0 n_2 (x_2 - x_0)^2$이 된다. 따라서 식 (2.4)에 제시된 최소제곱 기울기 추정량 $\hat{\beta}_{LS}$는 다음과 같이 표현된다.

$$\hat{\beta}_{LS} = \sum_{i<j} \frac{n_i n_j (x_j - x_i)^2}{n_0 n_1 (x_1 - x_0)^2 + n_1 n_2 (x_2 - x_1)^2 + n_0 n_2 (x_2 - x_0)^2} \hat{\beta}_{ij}. \tag{2.6}$$

구체적으로 x_0, x_1, x_2에 IBD 수인 0, 1, 2를 대입시켜 보면 가중의 합은 $n_0n_1 + n_1n_2 + 4n_0n_2$가 되며 따라서 $\hat{\beta}_{LS}$에서 $\hat{\beta}_{01}, \hat{\beta}_{12}, \hat{\beta}_{02}$ 각각의 가중은 $n_0n_1, n_1n_2, 4n_0n_2$가 된다. 전체 형제 쌍의 표본수가 클수록 두 형제의 IBD 수가 0, 1, 2가 되는 비율은 0.25, 0.5, 0.25에 근사하게 되므로 $n_0 = n_2 = m, n_1 = 2m$으로 가정해 본다. 이러한 가정 하에서 $\hat{\beta}_{01}, \hat{\beta}_{12}, \hat{\beta}_{02}$ 각각의 가중은 $2m^2, 2m^2, 4m^2$이 되어, $\hat{\beta}_{02}$가 나머지 두 경우를 합친 것과 동일한 비중으로 $\hat{\beta}_{LS}$에 기여하게 된다. 물론 이러한 비교에 $\hat{\beta}_{01}, \hat{\beta}_{12}, \hat{\beta}_{02}$ 각각의 기울기가 서로 다르게 추정됨이 감안되지 않았다.

이제 Haseman과 Elston (1972)의 기울기와 Kruglyak와 Lander (1995)의 통계량 X_{KL}을 비교해 본다. $n_0 = n_2 = m$의 가정 하에서 원 자료로 표현된 Kruglyak와 Lander (1995)의 통계량은 $m(\bar{Y_2} - \bar{Y_0})$이다. 여기서 $\bar{Y_2}$와 $\bar{Y_0}$는 IBD 수가 각각 2와 0인 값을 가지는 자료만의 평균이다. 한편 Haseman과 Elston (1972)의 '02' 부분군의 기울기는 $\hat{\beta}_{02} = (\bar{Y_2} - \bar{Y_0})/(x_2 - x_0) = (\bar{Y_2} - \bar{Y_0})/2$로서 결국 다음과 같은 관계가 성립한다.

$$\hat{\beta}_{02} = \frac{X_{KL}}{2m} \quad \text{또는} \quad X_{KL} = 2m\hat{\beta}_{02}. \tag{2.7}$$

따라서 두 통계량을 모두 원 자료에 근거하여 계산한다면 Kruglyak와 Lander (1995)의 통계량 X_{KL}은 Haseman과 Elston (1972)의 '02' 부분군의 기울기 $\hat{\beta}_{02}$와 상수 관계에 있음을 알 수 있고, 통계량 X_{KL}이 더욱 크다.

우리는 두 방법 모두 원 자료에 적용한 경우로 비교를 시작하였다. 다시 생각해 보아야 하는 점은 Kruglyak와 Lander (1995)의 통계량이 전체 순위 자료에 근거하면서도 IBD 수가 0과 2인 군만의 순위합의 함수로 표현하여 $X_{KL} = R_2 - R_0$인 것을 감안한다면 전체 자료에 근거한 Haseman과 Elston (1972)의 기울기 $\hat{\beta}_{LS}$와 Kruglyak와 Lander (1995)의 통계량을 비교함이 또 다른 의미를 가지게 된다. 우선 조건 $n_0 = n_2$하에서, 또한 x_0, x_1, x_2에 IBD 수인 0, 1, 2를 대입시켜 식 (2.6)의 기울기 추정량 $\hat{\beta}_{LS}$를 구하면 다음과 같다.

$$\hat{\beta}_{LS} = \frac{1}{2n_0(n_1 + 2n_0)}[n_0n_1(\hat{\beta}_{01} + \hat{\beta}_{12}) + 4{n_0}^2\hat{\beta}_{02}]. \tag{2.8}$$

이제 식 (2.5)의 $\hat{\beta}_{01}, \hat{\beta}_{12}, \hat{\beta}_{02}$를 대입한다.

$$\begin{aligned} \hat{\beta}_{LS} &= \frac{1}{2n_0(n_1 + 2n_0)}[n_0n_1(\bar{Y_2} - \bar{Y_0}) + 2{n_0}^2(\bar{Y_2} - \bar{Y_0})] \\ &= \frac{1}{2(n_1 + 2n_0)}[n_1(\bar{Y_2} - \bar{Y_0}) + 2n_0(\bar{Y_2} - \bar{Y_0})]. \end{aligned} \tag{2.9}$$

다시 말하면 두 방법은 다음과 같이 서로 연관된다. 전체 자료의 회귀식에서 $n_0 = n_2$의 조건하에서 구한 기울기 $\hat{\beta}_{LS}$는 바로 IBD 수가 2와 0인 군만의 평균차로 표현되며, 이는 바로 Kruglyak와 Lander (1995)의 통계량의 표현과 동일하다. 대략 n_1과 $2n_0$이 동일하고 식 (2.9)의 둘째 항이 $\hat{\beta}_{02}$에 기인한 부분이고 첫째 항이 $\hat{\beta}_{01}$과 $\hat{\beta}_{12}$에 기인한 부분임을 고려하면 $\hat{\beta}_{LS}$이 모두 $\bar{Y_2} - \bar{Y_0}$의 형태로만 표현되었다 하더라도 절반은 $\hat{\beta}_{01}$과 $\hat{\beta}_{12}$에 의한 기여로

서 결코 $\hat{\beta}_{02}$만의 기여가 아님을 알 수 있다. 이는 Kruglyak와 Lander (1995)의 통계량에서 전체 자료의 순위가 사용된 것과 같은 맥락이라 하겠다. 결론적으로 $n_0 = n_2$의 조건하에서 Kruglyak와 Lander (1995)의 통계량과 Haseman과 Elston (1972)의 통계량이 원 자료와 순위 자료의 차이만을 제외하고는 동일한 형태로 표현되지만 $n_0 \neq n_2$이 되면 두 통계량을 쉽게 연관 지을 수 없다.

2.2.2. 통계량의 분산

우선 Kruglyak와 Lander (1995)의 통계량 X_{KL}의 분산을 살펴본다. 이미 앞에서 언급한 바와 같이 유전 연관성이 없다는 가정 하에서 X_{KL}의 분산으로 Kruglyak와 Lander (1995)는 전체 표본수만의 함수인 $V = n(n+1)(2n+1)/12$을 제시하였다. 이러한 분산공식의 배경은 $n_0 = n_2 = m, n_1 = 2m$의 가정 하에서, 즉 IBD 수가 2와 0인 군의 자료가 전체 자료의 절반이 된다고 하면 X_{KL}의 기대값은 0이며 따라서 분산은 단순히 다음과 같다.

$$\begin{aligned} \mathrm{Var}(X_{KL}) = E({X_{KL}}^2) &= \frac{1}{2}\sum_{i=1}^{n} i^2 \\ &= \frac{1}{2}\frac{n(n+1)(2n+1)}{6} = \frac{n(n+1)(2n+1)}{12}. \end{aligned} \tag{2.10}$$

이는 기대되는 분산으로서 다시 말하면 수없이 반복된 실험 연구의 자료로부터 구해지는 평균적인 분산이다. 실제 수집된 자료에 근거한 분산을 다음과 같이 구할 수 있다.

Kruglyak와 Lander (1995) 통계량은 $X_{KL} = R_2 - R_0$이며, 여기서 R_2와 R_0은 IBD 수가 각각 2와 0인 군에 속한 자료의 순위합이며, 이 때 각 자료의 순위는 전체 자료의 순위이다. Hettmansperger (1984)이 유도한 순위합의 기대값과 분산, 즉, $E(R_i) = n_i(n+1)/2$, $\mathrm{Var}(R_i) = n_i(n-n_i)(n+1)/12$, $\mathrm{Cov}(R_i, R_j) = -n_i n_j(n+1)/12$ 을 이용하여 순위 통계량의 대표본 분산은 다음과 같다.

$$\mathrm{Var}(R_2 - R_0) = \frac{(n+1)[n(n_2+n_0) - (n_2-n_0)^2]}{12} \tag{2.11}$$

이 되며 $n_0 = n_2$의 가정 하에서는 $\mathrm{Var}(R_2 - R_0) = n(n+1)(2n_0)/12$으로 단순히 표현된다. Kruglyak와 Lander(1995)가 제시한 분산 공식인 $V = n(n+1)(2n+1)/12$과 비교해 본다. 추가적인 $n_0 = n_2 = m, n_1 = 2m$의 가정 하에서 식 (2.11)의 분산은 대략 $(4m)^2 \times 2m/12$인데 반해, Kruglyak와 Lander (1995)가 제시한 분산은 근사적으로 $(4m)^2 \times 8m/12$이므로 V가 대략 4배로 크다.

이제 원 자료에 근거한 회귀 직선식의 기울기 분산을 생각해 본다. 우리는 앞에서 Kruglyak와 Lander (1995)의 통계량이 $\hat{\beta}_{02}$, 그리고 $\hat{\beta}_{LS}$와 어떻게 관계되었는지 설명하였다. 따라서 $\hat{\beta}_{02}$와 $\hat{\beta}_{LS}$의 분산을 제시하며 두 분산의 크기를 또한 비교한다. 우선 부분군 $\hat{\beta}_{ij}$의 분산을 구하며 다음으로 전체 자료의 회귀식에서 $n_0 = n_2 = m$의 조건하에서 구한 회귀식의 기울기 $\hat{\beta}_{LS}$의 분산을 구해 본다.

이제 'ij' 부분군의 자료에 근거한 회귀 모형식 $Y_i = \alpha + \beta_{ij} x_l + e_l \, (l = n_i + n_j)$에서 $e_l \sim N(0, {\sigma_{ij}}^2)$를 가정한다. 여기서 ${\sigma_{ij}}^2$는 'ij' 부분군에 국한된, 그러나 IBD 수가 i인 군과

IBD 수가 j인 군의 공통분산이다. 한편 기울기 $\hat{\beta}_{ij}$의 최소제곱 추정량은 다음과 같다.

$$\begin{aligned}\hat{\beta}_{ij} = \frac{\bar{Y}_j - \bar{Y}_i}{x_j - x_i} &= \frac{\alpha + \beta_{ij}x_j + \bar{e_j} - (\alpha + \beta_{ij}x_i + \bar{e_i})}{x_j - x_i} \\ &= \beta_{ij} + \frac{\bar{e_j} - \bar{e_i}}{x_j - x_i}. \end{aligned} \tag{2.12}$$

여기서 β_{ij}는 참(true) 기울기로 상수이고, $\bar{e_j}$와 $\bar{e_i}$는 확률변수로 각각 n_j와 n_i개 오차의 평균이다. $\hat{\beta}_{ij}$의 대표본 근사 분산은

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_{ij}) &= \frac{1}{(x_j - x_i)^2}\mathrm{Var}(\bar{e_j} - \bar{e_i}) \\ &= \left(\frac{1}{n_j} + \frac{1}{n_i}\right){\sigma_{ij}}^2/(x_j - x_i)^2 \end{aligned} \tag{2.13}$$

이며 ${\sigma_{ij}}^2$의 추정량으로서 'ij' 부분군의 회귀 모형식에 의한 MSE_{ij}(Mean squares of error of the subset 'ij')를 사용하여 $\mathrm{Var}(\hat{\beta}_{ij})$의 추정량을 구할 수 있다. 이제 '$ij$'부분군의 자료로부터 $\bar{Y} = (n_i\bar{Y_i} + n_j\bar{Y_j})/(n_i + n_j)$, $\bar{x} = (n_i\bar{x_i} + n_j\bar{x_j})(n_i + n_j)$가 구해지고 이로부터 $\hat{\alpha} = \bar{Y} - \hat{\beta}_{ij}\bar{x}$와 식 (2.5)에 제시된 $\hat{\beta}_{ij}$을 이용하여 추정값 $\hat{\alpha} + \hat{\beta}_{ij}x_i$는 단순히 $\bar{Y_i}$가 되고, $\hat{\alpha} + \hat{\beta}_{ij}x_j$는 $\bar{Y_j}$가 되며, 이는 상식적으로 쉽게 이해된다. 따라서 ${\sigma_{ij}}^2$의 추정량으로서 단지 두 군으로 구성된 'ij' 부분군의 회귀 직선식의 MSE_{ij}인 ${s_{ij}}^2$는 합동 표본분산(pooled sample variance)으로

$${s_{ij}}^2 = \frac{\sum^{n_i}(Y_{l_i} - \bar{Y_i})^2 + \sum^{n_j}(Y_{l_j} - \bar{Y_j})^2}{n_i + n_j - 2} \tag{2.14}$$

이며 이를 식 (2.13)에 대입하여 $\hat{\mathrm{Var}}(\hat{\beta}_{ij})$을 구한다. 여기서 x_0, x_1, x_2가 지정되어 있지 않을 때 각 IBD 군 자료의 합산을 단순히 $1/n_i\sum^{n_i} Y_{l_i}$로 표기하였다.

이제 전체 자료의 회귀식에서 $n_0 = n_2$의 조건하에서 구한 회귀식의 기울기 $\hat{\beta}_{LS}$의 분산을 구해 본다. $\hat{\beta}_{LS}$의 분산에 각 부분군에 근거한 $\hat{\beta}_{02}$의 분산과 비교하기 위해 식 (2.6)의 표현으로부터 출발한다. 각 부분군의 참 분산은 모두 σ^2로 가정한다. 상수 $C = n_0n_1 + n_1n_2 + 4n_0n_2$, $C_0 = n_0n_1/C$, $C_1 = n_1n_2/C$, $C_2 = 4n_0n_2/C$를 정의한다. $\hat{\beta}_{LS} = C_0\hat{\beta}_{01} + C_1\hat{\beta}_{12} + C_2\hat{\beta}_{02}$에서 각 부분군의 분산 식 (2.13)과 또한 $\mathrm{Cov}(\hat{\beta}_{01}, \hat{\beta}_{12}) = -\sigma^2/n_1$, $\mathrm{Cov}(\hat{\beta}_{01}, \hat{\beta}_{02}) = \mathrm{Cov}(\hat{\beta}_{12}, \hat{\beta}_{02}) = \sigma^2/(2n_0)$을 다음 식에 대입하여 $\hat{\beta}_{LS}$의 분산을 구한다.

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_{LS}) &= {C_0}^2\mathrm{Var}(\hat{\beta}_{01}) + {C_1}^2\mathrm{Var}(\hat{\beta}_{12}) + {C_2}^2\mathrm{Var}(\hat{\beta}_{02}) \\ &\quad +2C_0C_1\mathrm{Cov}(\hat{\beta}_{01}, \hat{\beta}_{12}) + 2C_0C_2\mathrm{Cov}(\hat{\beta}_{01}, \hat{\beta}_{02}) \\ &\quad +2C_1C_2\mathrm{Cov}(\hat{\beta}_{12}, \hat{\beta}_{02}). \end{aligned} \tag{2.15}$$

$\hat{\beta}_{LS}$의 분산 공식 (2.15)에서 '02' 부분군의 분산 부분, 즉 오른쪽 세 번째 항과 나머지 항들을 비교해 보면 '02' 부분군의 분산 부분은 $2n_0\sigma^2/n^2$이고 나머지 항들의 분산 부분은 $n_1(n_1 + 4n_0)\sigma^2/(2n_0n^2)$이다. 따라서 전자에 대한 후자의 비는 $n_1(n_1 + 4n_0)/(4{n_0}^2)$이 되며

$n_0 = n_2 = m, n_1 = 2m$의 가정 하에서 3이 된다. 이로부터 $\hat{\beta}_{LS}$의 분산과 부분군 $\hat{\beta}_{02}$의 분산이 $\mathrm{Var}(\hat{\beta}_{LS})/\mathrm{Var}(\hat{\beta}_{02}) = 4$의 관계에 있음을 알 수 있다. 이는 $\hat{\beta}_{02}$에 근거한 검정의 p값이 $\hat{\beta}_{LS}$에 근거한 검정의 p값보다 더욱 작을 수 있는 가능성을 말해 준다. 그러나 기울기 및 기울기의 분산, 그리고 자유도의 총체적인 결과로 검정력이 결정되기 때문에 분산의 크기만으로 검정력을 단정 지을 수는 없다. 이제 종합적인 통계량의 비교를 다음 절에서 제시한다.

2.2.3. 검정 통계량의 비교

Kruglyak와 Lander (1995)의 통계량은 순위 자료에 근거한 $X_{KL} = R_2 - R_0$이며 $Z_{KL} = X_{KL}/\sqrt{V}$이 근사적으로 표준 정규분포를 따름을 이용하여 검정한다. 이에 대응하는 '02' 부분군의 기울기 β_{02}가 0인 귀무가설의 검정통계량을 살펴보면 다음과 같다.

$$T = \frac{\hat{\beta}_{02}}{\sqrt{\hat{\mathrm{Var}}(\hat{\beta}_{02})}} = \frac{\bar{Y_2} - \bar{Y_0}}{\sqrt{\frac{n_0+n_2}{n_0 n_2 (n_0+n_2-2)}[\sum^{n_0}(Y_{l_0} - \bar{Y_0})^2 + \sum^{n_2}(Y_{l_2} - \bar{Y_2})^2]}}. \tag{2.16}$$

음수로 기울기가 클 때 T가 자유도 $n_0 + n_2 - 2$인 t 분포함을 이용하여 단측 검정한다. 그러나 Haseman과 Elston (1972)의 회귀식은 전체 자료에 근거한다. 기울기 β_{LS}가 0인 가설의 검정 통계량은 $\hat{\beta}_{LS}/\sqrt{\mathrm{Var}(\hat{\beta}_{LS})}$이며 이 과정에서 요구되는 MSE는 전체 자료의 회귀식으로부터 구한 잔차의 제곱합을 $n-2$로 나눈 것이다. 이 $\hat{\beta}_{LS}/\sqrt{\mathrm{Var}(\hat{\beta}_{LS})}$을 자유도 $n-2$의 t 분포함을 이용하여 단측 검정한다. 이는 '02' 부분군의 경우인 식 (2.16)과 같이 간단한 형태로 요약되지 않는다.

순위 자료에 근거한 Kruglyak와 Lander (1995)의 통계량 Z_{KL}과 원 자료에 근거한 Haseman과 Elston (1972)의 통계량은 근거하는 자료, 통계량 및 자유도가 다르기 때문에 즉각적인 비교가 어렵다. 두 군의 순위합에 근거한 Kruglyak와 Lander (1995)의 통계량이 오히려 전체 자료를 사용하는 Haseman과 Elston (1972)의 통계량보다 더욱 유의한 경우도 있겠다. 따라서 모의실험을 이용한 비교가 절실히 요구되며 각 통계량에서 요구되는 가정과 함께 모의실험으로 검정력을 비교할 것이다.

2.3. 가중 회귀분석법

Haseman과 Elston (1972) 회귀검정의 가정 중 하나는 오차항의 분산이 σ^2으로 일정하다는 것이다. 일반적으로 알려져 있지 않은 분산의 추정량으로 적합된 회귀식의 잔차 제곱합의 평균을 사용한다. 그러나 설명변수의 값에 자료가 반복되어 있고, 더욱이 반복수가 다르기 때문에 등분산 가정이 성립하지 않는다. 또한 유전연관성 검정에서 정규분포를 따르는 형질의 형제 쌍 차의 제곱의 분산은 IBD 수가 0, 1, 2로 변해가면서 감소 경향을 보인다고 앞에서 언급하였다. 다시 말해서 이분산성(variance heterogeneity) 하에서 적합한 Haseman과 Elston (1972) 회귀검정을 제안하며 특히 종속변수를 가중(weight)시켜 회귀분석하는 방법을 제시한다. 종속변수의 변수변환도 분산을 안정화 시킬 수 있지만 이는 기울기의 추정량이 최소분산 특성을 갖지 않으며 추정값과 예측값이 변환된 척도로 얻어지므

로 해석이 용이하지 않다. 따라서 설정된 회귀모형이 적절하고 오차항의 분산만 동일하지 않다면 가중 최소제곱법(method of weighted least squares)이 바람직하다. 일반적으로 종속변수의 역분산(inverse variance)을 가중값으로 선택한다. 만약 분산이 알려져 있다면 가중값은 상수이지만 실제로 분산을 모르기 때문에 자료 분석시에 분산을 추정하여 가중값을 정하게 된다. 특히 각 IBD 수에 연속 형질의 자료가 과도 반복된 유전연관성 검정에서 가중 최소제곱 회귀검정은 매우 적절한데 그 이유는 분산을 추정하기 위해서는 반복수가 커야 하기 때문이다. 반복수가 매우 적으면($n_i < 10$) 가중 최소제곱 추정량은 일관성이 없으며 자료의 분포를 나타내는 추정량으로 적합하지 않다 (Carroll과 Cline, 1988).

2.3.1. 가중 회귀식의 기울기

반복이 있는 이분산 회귀모형은 다음과 같다.

$$Y_{ij} = \alpha + \beta x_i + e_{ij}, \qquad j = 1, \ldots, n_i;\ n_i > 1. \tag{2.17}$$

각 x_i에 반복수는 n_i이며 e_{ij}는 평균이 0이고 분산은 ${\sigma_i}^2$인 정규분포를 따르며 각각 독립임을 가정한다. 각 x_i마다 서로 다른 ${\sigma_i}^2$로 인해 이분산성이 대두된다. ${\sigma_i}^2$의 추정량으로 ${s_i}^2 = \sum_j (y_{ij} - \bar{y_i})^2/(n_i - 1)$를 제시할 때 가중 회귀식은

$$Q = \sum_i \sum_j (y_{ij} - \alpha - \beta x_i)^2 / {s_i}^2 \tag{2.18}$$

을 최소로 하는 절편과 기울기의 추정량 a와 b를 구하게 되며, 여기서 $1/{s_i}^2$가 추정된 가중이다. 반복이 있는 회귀모형에서 $\hat{\omega}_i = n_i/{s_i}^2$로 정의하면 a와 b는 다음과 같다.

$$\begin{aligned} a &= \frac{\sum \hat{\omega}_i \bar{Y_i} \sum \hat{\omega}_i {x_i}^2 - \sum \hat{\omega}_i x_i \sum \hat{\omega}_i x_i \bar{Y_i}}{\sum \hat{\omega}_i \sum \hat{\omega}_i {x_i}^2 - (\sum \hat{\omega}_i x_i)^2}, \\ b &= \frac{\sum \hat{\omega}_i \sum \hat{\omega}_i x_i \bar{Y_i} - \sum \hat{\omega}_i x_i \sum \hat{\omega}_i \bar{Y_i}}{\sum \hat{\omega}_i \sum \hat{\omega}_i {x_i}^2 - (\sum \hat{\omega}_i x_i)^2}. \end{aligned} \tag{2.19}$$

일반적으로 가중 회귀식에서 가중이 추정되어야 함에도 불구하고 자료로부터 추정된 후에는, 가중을 주어진 경우로 취급하여 기울기의 분산을 구한다. Jacquez 등 (1968)은 가중 $\hat{\omega}_i$이 확률변수임을 감안하여 식 (2.19)에 제시된 기울기 b의 분산을 다음과 같이 제시하였다.

$$\begin{aligned} {s_b}^2 &= \frac{\sum \hat{\omega}_i}{\hat{S}} \left\{ 1 - 4 \sum_j \frac{1}{n_j - 1} \left[1 + \frac{\hat{D_j}}{\hat{S}} \right] \left[\frac{\hat{\omega}_j}{\sum \hat{\omega}_i} + \frac{\hat{D_j}}{\hat{S}} \right] \right\}, \\ \hat{D_j} &= 2\hat{\omega}_j x_j \sum \hat{\omega}_i x_i - \hat{\omega}_j \sum \hat{\omega}_i {x_i}^2 - \hat{\omega}_j {x_j}^2 \sum \hat{\omega}_i, \\ \hat{S} &= \sum \hat{\omega}_i \sum \hat{\omega}_i {x_i}^2 - \left(\sum \hat{\omega}_i x_i \right)^2. \end{aligned} \tag{2.20}$$

우리는 각 x_i에서 반복수와 분산이 다름을 감안한 가중값 $\hat{\omega}_i = n_i/{s_i}^2$을 사용한 가중 최소제곱법의 기울기와 Jacquez 등 (1968)이 유도한 분산으로 유전연관성 검정을 시행할 것을 제안하며 본 논문에서 언급한 Haseman과 Elston (1972) 회귀검정과 Kruglyak와 Lander (1995)의 통계량을 이용한 검정결과와 비교하고자 한다. 또한 연속 형질의 자료가 정규분포 또는 비정규분포를 따름에 따라 각각 어떤 검정법이 더 효과적인지 모의실험을 통하여 알아보겠다.

3. 모의실험

3.1. 모의실험의 계획

모의실험 자료는 다음과 같이 생성한다. j번째 형제 쌍의 첫째 형제와 둘째 형제의 형질을 각각 Y_{1j}와 Y_{2j}라 할 때 Y_{1j}와 Y_{2j}의 일반적인 모형을 다음과 같이 가정한다.

$$\begin{aligned} Y_{1j} &= \mu + g_{1j} + \epsilon_{1j}, \\ Y_{2j} &= \mu + g_{2j} + \epsilon_{2j}. \end{aligned} \tag{3.1}$$

여기서 μ는 전체평균이고, g_{ij}는 유전적 효과(genetic effect), 그리고 ϵ_{ij}는 환경적 효과 (environmental effect)를 나타낸다. 대립유전자 B와 b가 있을 때 각각의 대립유전자 빈도(allele frequency)는 p와 $q(=1-p)$로 나타내며, 모의실험에서 p는 0.1, 0.3, 0.5, 0.7, 0.9로 가정한다. 유전자형에 따라 변하는 g_{ij}는 BB의 경우에 a이고, bb의 경우에 $-a$, 그리고 Bb의 경우에 d로 정의한다. a와 d의 결정은 곧 소개하게 되는 유전율에 따라 정한다. 환경적 효과를 나타내는 ϵ_{ij}는 모수적 방법과 비모수적 방법의 효율성을 비교하기 위해 정규분포 또는 비정규분포로 생성하였다. 정규분포의 경우 평균이 0이고 분산이 1인 표준 정규분포로 생성하였고, 비정규분포로서 대칭적인 혼합된(contaminated) 정규분포와 치우친(skewed) 분포로서 로그정규분포로 생성하였다. 혼합된 정규분포의 누적함수는 구체적으로 $(1-\delta)\Phi(x)+\delta\Phi(x/k)$에서 $\delta = 0.025$와 $k = 5$로 두어 전체 표준 정규분포 자료의 2.5%는 평균이 0이고 표준편차가 5인 정규분포로 대치하였다. 한편 두 형제의 연속 형질의 상관성은 0.5가 되도록 생성하였다. 모의실험에서 비정규분포 자료를 고려한 이유는 의학에서 정규분포를 따르는 연속 형질 자료도 있지만 정규분포가 아닌 자료도 있기 때문이다. 예를 들어 중성지방, 혈색소, 혈소판 수, 백혈구 수 등은 정규분포하를 따르는 것으로 알려져 있지만, 콜레스테롤과 LDL-콜레스테롤 수치는 정규분포를 따르지 않는다 (이진찬 등, 2006).

형질의 표현형이 다음 세대에 어느 정도 유전되는지를 나타내는 유전율(heritability) h^2는 다음과 같이 정의된다.

$$h^2 = \frac{{\sigma_g}^2}{{\sigma_g}^2 + {\sigma_e}^2}. \tag{3.2}$$

본 논문에서 h^2는 각각 0, 0.2, 0.4, 0.6, 0.8로 가정한다. 한편 ${\sigma_e}^2 = 1$로 가정하고 유전자 효과 g_{ij}의 분산인 ${\sigma_g}^2$는 ${\sigma_a}^2 + {\sigma_d}^2$로 표현되며 ${\sigma_a}^2$과 ${\sigma_d}^2$는 각각 다음과 같이 주어진다

(Haseman과 Elston, 1972).

$$\begin{aligned} {\sigma_a}^2 &= 2pq[a-d(p-q)]^2, \\ {\sigma_d}^2 &= 4p^2q^2d^2. \end{aligned} \tag{3.3}$$

우성 모형(dominance model)의 경우에 $d = a$이고 가법 모형(additive model)인 경우에 $d = 0$이며, 열성 모형(recessive mode)이 되면 $d = -a$이다. 따라서 유전율과 대립유전자 빈도가 정해지면 식 (3.2)와 (3.3)에 의해 각 모형 하에서 a와 d가 계산된다.

자료 생성의 다음 단계로 두 형제가 대립유전자를 공유하는 정도, 즉, IBD 수에 따라 두 형제의 유전자형을 생성한다. 우선 두 형제의 형질 IBD 수가 각각 0, 1, 2가 되는 비율 0.25, 0.5, 0.25가 되도록 다항분포(multinomial distribution) 자료를 랜덤하게 생성한다. 각 개체의 유전자형의 BB, Bb, bb 중 하나로 각각의 발생확률은 $p^2, 2pq, q^2$이다 (Falconer와 Mackay, 1996). 따라서 두 형제의 유전자형의 조합으로서 9가지가 있으며, 앞에서 생성된 IBD 수에 따라 각 경우의 조건부 확률은 Haseman과 Elston (1972)의 논문에 수록되어 있는 표 1에 근거한다 (편의상 표 1을 본 논문에 표 3.1로 제시하였다). 표 3.1을 간략히 설명하면 두번째 열에는 두 형제의 형질 차의 제곱인 $Y_j = (Y_{1j} - Y_{2j})^2$이 제시되었으며 이는 단순히 두 형제의 유전자형 조합에 따라 식 (3.1)로부터 구해진 것이다. 예를 들어, 두 형제의 유전자형 조합이 $BB - Bb$이면 $x_{1j} = \mu + a + e_{1j}$와 $x_{2j} = \mu + d + e_{2j}$로부터 $Y_j = (a + e_{1j} - d - e_{2j})^2 = (a - d + e_j)^2$이 된다.

표 3.1의 마지막 세 열에 제시된 π_j에 따른 두 형제 유전자형 조합의 조건부 확률은 다음과 같이 구한다. $\pi_j = 0$인 경우는 두 형제가 공유하는 대립유전자가 없는 서로 독립된 상황이므로 해당하며 이러한 경우의 확률은 각 형제의 유전자형의 확률을 곱한 것이 된다. $\pi_j = 1$인 경우는 두 형제가 서로 동일한 유전자형을 가진 경우로 처음 세 행만이 가능하여 각 행의 해당 확률은 한 형제의 유전자형의 확률인 $p^2, 2pq, q^2$으로 결정된다. $\pi_j = 0.5$는 두 형제가 한 개의 대립유전자만을 공유로 하는 경우로 공유되는 대립유전자에 대한 확률과 그렇지 않은 대립유전자를 가질 확률의 곱으로 구해진다. 즉, $BB - BB$이면 p^3이고, $bb - bb$이면 q^3이다. $BB - Bb$이면 $p \times p \times q$가 되어 p^2q이고, $Bb - bb$이면 $q \times q \times p$가 되어

표 3.1: π_j에 따른 두 형제 유전자형의 조건부 조합 확률

		Conditional probability		
Sib pair	Y_j	$\pi_j = 0$	$\pi_j = \frac{1}{2}$	$\pi_j = 1$
$BB - BB$	${e_j}^2$	p^4	p^3	p^2
$bb - bb$	${e_j}^2$	q^4	q^3	q^2
$Bb - Bb$	${e_j}^2$	$4p^2q^2$	pq	$2pq$
$BB - Bb$	$(a - d + e_j)^2$	$2p^3q$	p^2q	0
$Bb - BB$	$(-a + d + e_j)^2$	$2p^3q$	p^2q	0
$Bb - bb$	$(a + d + e_j)^2$	$2pq^3$	pq^2	0
$bb - Bb$	$(-a - d + e_j)^2$	$2pq^3$	pq^2	0
$BB - bb$	$(2a + e_j)^2$	p^2q^2	0	0
$bb - BB$	$(-2a + e_j)^2$	p^2q^2	0	0

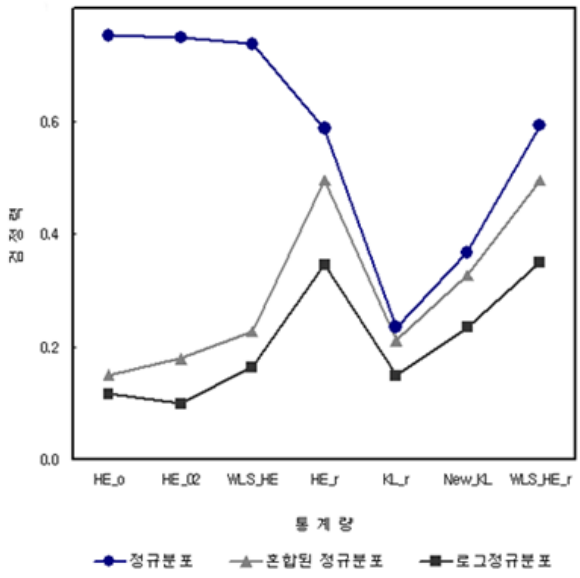

그림 3.1: 분포에 따른 여러 검정법의 검정력

pq^2이다. 마지막으로 $Bb - Bb$이면 $pq^2 + p^2q = pq$이다. 이제 표 3.1의 조건부 확률을 이용하여 두 형제의 유전자형의 조합을 생성하고 앞에서 언급한 방법들의 검정력을 비교한다.

본 논문에서 언급한 여러 검정법들의 비교로서 다음의 검정법들을 모의실험에서 비교할 것이다. 원 자료에 Haseman과 Elston (1972) 회귀검정(HE)을 적용하고, '02' 부분군의 자료에 Haseman과 Elston (1972) 회귀검정(HE_02)을 적용한다. 이는 $\hat{\beta}_{02}$에 근거한 검정의 p값이 $\hat{\beta}_{LS}$에 근거한 검정의 p값보다 더욱 작을 수 있는 가능성을 확인해 보기 위함이다. 이제 전체 자료를 순위 자료로 변환하여 Haseman과 Elston (1972) 회귀검정(HE_r)을 적용하고, 순위 자료를 Kruglyak와 Lander (1995)의 방법으로 검정(KL)한다. 식 (2.11)에 제시한 분산을 이용하여 Kruglyak와 Lander (1995)의 방법을 검정(KL_n)하여 조정된 분산일 경우의 검정력과 비교한다. 마지막으로 원 자료에 가중 회귀검정(WLS_HE)을, 순위 자료에 가중 회귀검정(WLS_HE_r)을 적용한다. 이는 일부 검정법은 원 자료에 근거하고 일부 검정법은 순위 자료에 근거하기 때문이다.

가법 모형 하에서 모의실험을 진행하였으며, 표본의 크기는 여러 문헌에 근거하여 500개의 형제 쌍 자료를 생성하였고 모의실험의 반복회수는 1,000번으로 정하였다.

3.2. 모의실험의 결과

표 3.2는 정규분포 하에서, 표 3.4는 대칭적인 혼합된 정규분포 하에서 서로 다른 대립유전자 빈도에 따라 여러 검정법의 유의수준을 제시한 것이다. 정규분포 하에서 유의수준은

표 3.2: 정규분포를 따르는 형질 자료에 적용한 여러 검정법의 유의수준

		original data			rank transformed data			
p	h^2	HE	HE_02	WLS_HE	HE_r	KL	KL_n	WLS_HE_r
0.1	0	0.043	0.045	0.052	0.051	0.042	0.048	0.049
0.3	0	0.052	0.047	0.050	0.038	0.043	0.040	0.036
0.5	0	0.046	0.047	0.049	0.045	0.049	0.042	0.048
0.7	0	0.051	0.047	0.055	0.056	0.049	0.044	0.057
0.9	0	0.050	0.047	0.049	0.056	0.052	0.045	0.054

표 3.3: 정규분포를 따르는 형질 자료에 적용한 여러 검정법의 검정력

		original data			rank transformed data			
p	h^2	HE	HE_02	WLS_HE	HE_r	KL	KL_n	WLS_HE_r
0.1	0.2	0.705	0.696	0.705	0.460	0.191	0.307	0.464
	0.4	0.996	0.997	0.997	0.905	0.444	0.621	0.909
	0.6	1.000	1.000	1.000	0.993	0.607	0.800	0.995
	0.8	1.000	1.000	1.000	0.999	0.686	0.859	0.999
0.3	0.2	0.692	0.707	0.706	0.542	0.195	0.344	0.550
	0.4	0.998	0.999	0.999	0.979	0.579	0.787	0.980
	0.6	1.000	1.000	1.000	1.000	0.901	0.977	1.000
	0.8	1.000	1.000	1.000	1.000	0.986	0.997	1.000
0.5	0.2	0.752	0.748	0.737	0.588	0.235	0.367	0.593
	0.4	0.999	0.999	0.999	0.987	0.585	0.794	0.984
	0.6	1.000	1.000	1.000	1.000	0.937	0.979	1.000
	0.8	1.000	1.000	1.000	1.000	0.998	1.000	1.000
0.7	0.2	0.732	0.735	0.739	0.536	0.216	0.330	0.544
	0.4	0.997	0.998	0.998	0.981	0.564	0.774	0.983
	0.6	1.000	1.000	1.000	1.000	0.886	0.966	1.000
	0.8	1.000	1.000	1.000	1.000	0.995	1.000	1.000
0.9	0.2	0.710	0.722	0.708	0.476	0.194	0.301	0.476
	0.4	0.996	0.995	1.000	0.920	0.426	0.654	0.923
	0.6	1.000	1.000	1.000	0.990	0.645	0.838	0.992
	0.8	1.000	1.000	1.000	0.997	0.694	0.863	0.999

표 3.4: 비정규분포하는 형질 자료에 적용한 여러 검정법의 유의수준

		original data			rank transformed data			
p	h^2	HE	HE_02	WLS_HE	HE_r	KL	KL_n	WLS_HE_r
0.1	0	0.048	0.038	0.077	0.046	0.050	0.045	0.045
0.3	0	0.052	0.042	0.098	0.053	0.053	0.051	0.051
0.5	0	0.059	0.054	0.115	0.047	0.051	0.049	0.053
0.7	0	0.042	0.049	0.086	0.048	0.048	0.054	0.049
0.9	0	0.060	0.046	0.093	0.050	0.052	0.047	0.052

표 3.5: 비정규분포하는 형질 자료에 적용한 여러 검정법의 검정력

		original data			rank transformed data			
p	h^2	HE	HE_02	WLS_HE	HE_r	KL	KL_n	WLS_HE_r
0.1	0.2	0.151	0.182	0.219	0.457	0.186	0.293	0.455
	0.4	0.431	0.489	0.485	0.875	0.384	0.612	0.880
	0.6	0.829	0.825	0.833	0.983	0.576	0.786	0.988
	0.8	0.997	0.997	0.997	0.995	0.662	0.831	0.996
0.3	0.2	0.128	0.163	0.207	0.501	0.209	0.318	0.510
	0.4	0.398	0.466	0.464	0.956	0.518	0.723	0.959
	0.6	0.856	0.841	0.869	1.000	0.870	0.957	1.000
	0.8	1.000	1.000	1.000	1.000	0.980	0.996	1.000
0.5	0.2	0.152	0.180	0.227	0.496	0.213	0.327	0.496
	0.4	0.413	0.495	0.498	0.964	0.530	0.736	0.967
	0.6	0.842	0.828	0.859	1.000	0.867	0.960	1.000
	0.8	0.999	0.999	0.999	1.000	0.992	1.000	1.000
0.7	0.2	0.125	0.167	0.208	0.516	0.169	0.306	0.516
	0.4	0.431	0.479	0.478	0.963	0.511	0.751	0.968
	0.6	0.848	0.861	0.879	1.000	0.875	0.962	1.000
	0.8	1.000	0.998	1.000	1.000	0.978	0.993	1.000
0.9	0.2	0.150	0.175	0.216	0.454	0.175	0.298	0.457
	0.4	0.430	0.487	0.488	0.876	0.398	0.613	0.884
	0.6	0.814	0.835	0.847	0.978	0.597	0.781	0.986
	0.8	0.999	0.999	0.999	0.988	0.673	0.842	0.992

대부분 0.05 안팎으로 비슷하였으나 대칭적인 혼합된 정규분포 하에서의 결과인 표 3.4를 보면 순위 자료의 추정된 유의수준이 0.05 안팎인 것에 반하여 원 자료의 추정된 유의수준은 그보다 약간 높았다. 치우친 분포로서 로그정규분포 하에서의 검정력은 대칭적인 혼합된 정규분포 하에서의 경우보다 약간 떨어지지만 비슷한 경향을 보이고 있어 표로 따로 제시하지는 않고 $p = 0.5$이고 $h^2 = 0.2$인 경우의 모의실험 검정력 결과를 그림 3.1에 제시하였다.

표 3.3은 정규분포 하에서, 표 3.5는 비정규분포 하에서 유전율이 0.2, 0.4, 0.6, 0.8이고 대립유전자 빈도가 0.1, 0.3, 0.5, 0.7, 0.9로 다를 때 여러 검정법의 검정력 결과이다. 공통적인 결과는 가법 모형에서 대립유전자 빈도에 따른 검정력은 서로 비슷한 경향을 나타내었으며, 대립유전자 빈도가 같을 때 유전율이 증가할수록 검정력이 높게 나타났다. 대부분의 상황에서 Haseman과 Elston (1972) 회귀검정보다 서로 다른 반복수를 가중으로서 보정해 준 가중 회귀분석법의 검정력이 더 높았다. 최소제곱법의 검정력이 더 낮은 경우는 유전율이 0.2인 경우이지만 표에서 볼 수 있듯이 항상 그런 것은 아니다. '02' 부분군에서 Haseman과 Elston (1972) 회귀검정의 검정력이 원 자료에서 Haseman과 Elston (1972) 회귀검정보다 더 높은 경우가 있다. 2.2.2절에서 설명하였듯이 $\hat{\beta}_{02}$의 분산이 $\hat{\beta}_{LS}$의 분산의 1/4 정도이므

로 이러한 결과가 발생하였다. 물론 자유도가 약 1/2로 작아지는 것도 간과할 수 없으나 모의실험에서 표본수가 500쌍이었으므로 자유도에 의한 차이는 거의 없다고 볼 수 있다. 그러나 표본수가 작게 되면 다를 것으로 추측된다. 식 (2.11)의 조정된 분산을 사용한 새로이 제시된 방법은 Kruglyak와 Lander (1995)의 방법보다 항상 검정력이 더 높게 나왔다. 즉 조정된 분산은 기존의 분산보다 1/4 정도이기 때문에 대립가설 하에서 기각이 더 빈번함을 알 수 있다. 또한 비모수적 방법인 Kruglyak와 Lander (1995)의 방법보다 순위자료에서 가중 회귀분석법이 더 높은 검정력을 나타내어 비정규분포 자료에서도 가중 회귀분석법이 적절함을 알 수 있다. 여러 검정법들의 우성과 열성 모형 하에서 검정력의 우열은 변함이 없었다.

여러 방법 간의 결과를 요약하면 표 3.3에서 원 자료와 순위 자료의 결과를 보면 원 자료의 검정력이 더 높게 나타났다. 반면에 표 3.5에서는 순위 자료의 검정력이 더 높게 나타났다. Kruglyak와 Lander (1995)의 방법이 가장 낮은 검정력을 나타내었고, Haseman과 Elston (1972) 회귀검정보다 가중 회귀분석법이 높은 검정력을 나타내었다. 비정규분포 자료의 경우에는 순위 자료로 변환하여 검정하는 것이 더 높은 검정력을 나타내었다.

4. 결론 및 고찰

유전연관성 연구는 질병을 가지고 있는 가계나 형제로부터 자료를 수집하여 질병과 함께 부모로부터 전달되는 마커의 염색체상 위치를 확인한 후, 보다 세밀하게 질병의 원인으로 짐작되는 특정 유전자와 유전적 변형의 연관성을 규명하여 질병의 발생 기전을 밝히는 방법이다. 이 방법은 단일유전자로 인한 질병뿐만 아니라 여러 유전자의 복합적인 원인에 의한 질병의 유전자 탐색에도 유용하게 활용된다.

이 논문에서는 이러한 단일 유전자로 인한 유전연관성 검정의 방법으로 Kruglyak와 Lander (1995)의 방법과 Haseman과 Elston (1972) 회귀검정을 비교해 보았다. 한정된 설명변수의 값에 매우 많은 자료가 반복되어 있는 점을 보완한 가중 회귀분석법을 제안하였으며 검정력은 모의실험으로 알아보았다. 원 자료가 정규분포를 따르는 연속 형질의 경우에는 정규성 가정을 만족해야 되는 Haseman과 Elston (1972) 회귀검정이나 가중 회귀분석법이 적절하다. 반면에 원 자료가 비정규분포를 따르는 연속 형질의 경우에는 원 자료를 순위자료로 변환하여 가중 회귀분석법을 하는 것이 더욱 적절하다. 가중 회귀분석법은 Jacquez 등 (1968)이 제안한 가중 최소제곱 추정량을 사용하여 분석했는데 검정력은 가장 높았으나 계산과정이 다소 복잡한 단점이 있다. 비정규분포 자료라 하더라도 Kruglyak와 Lander (1995)의 방법이 비모수적 방법으로는 적절하지 않음을 알 수 있다.

참고문헌

강근석, 김충락 (2001). <회귀분석>, 교우사.

이진찬, 김순기, 이창규, 이승관, 이현실, 조경진 (2006). 건강검진 결과를 이용한 한국 성인 참고치 설정과 해석에 관한 고찰, <임상검사와 정도관리>, **28**, 229–237.

Carroll, R. J. and Cline, D. B. H. (1988). An asymptotic theory for weighted least-squares with weights estimated by replication, *Biometrika*, **75**, 35–43.

Dietz, E. J. (1989). Teaching regression in a nonparametric statistics course, *The American Statistician*, **43**, 35–40.

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

Hettmansperger, T. P. (1984). *Statictical Inference Based on Ranks,* John Wiley & Sons, New York.

Jacquez, J. A., Mather, F. J. and Crawford, C. R. (1968). Linear regression with nonconstant, unknown error variances: Sampling experiments with least squares, weighted least squares and maximum likelihood estimators, *Biometrics,* **24,** 607–626.

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.

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

Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis, *Journal of the American Statistical Association,* **47,** 583–621.

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

[2007년 7월 접수, 2007년 9월 채택]

Comparisons of Kruglyak and Lander's Nonparametric Linkage Test and Weighted Regression Incorporating Replications

Eun-Kyeong Choi[1] Hae-Hiang Song[2]

Abstract

The ordinary least squares regression method of Haseman and Elston(1972) is most widely used in genetic linkage studies for continuous traits of sib pairs. Kruglyak and Lander(1995) suggested a statistic which appears to be a nonparametric counterpart to the Haseman and Elston(1972)'s regression method, but in fact these two methods are quite different. In this paper the relationships between these two methods are described and will be compared by simulation studies. One of the characteristics of the sib-pair linkage study is that the explanatory variable has only three different values and thus dependent variable is heavily replicated in each value of the explanatory variable. We propose a weighted least squares regression method which is more appropriate to this situation and the efficiency of the weighted regression in genetic linkage study was explored with normal and non-normal simulated continuous traits data. Simulation studies demonstrated that the weighted regression is more powerful than other tests.

Keywords: Haseman and Elston regression, Kruglyak and Lander nonparametric statistic, sib pairs, linkage test.

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