한국보건정보통계학회지 제34권 제1호
Journal of the Korean Society of Health Information and Health Statistics
Volume 34, Number 1, 2009, pp. 53～62

양적 형질의 유전적 관련성 연구에서 치료 효과를 보정하기 위한 통계학적 방법의 비교

한경화[1], 임길섭[2], 박성하[3], 장양수[4], 송기준[2)†]
[1]연세대학교 의과대학 연구부, [2]연세대학교 의과대학 의학통계학과,
[3]연세대학교 의과대학 내과학교실, [4]연세대학교 의과대학 심혈관계질환 유전체연구센터

A comparison of statistical methods for adjusting the treatment effects in genetic association studies of quantitative traits

Kyung Hwa Han[1], Kil Seob Lim[2], Sung Ha Park[3],Yang Soo Jang[4], Ki jun Song[2)†]
[1]Department of Research Affairs, Yonsei University College of Medicine
[2]Department of Biostatistics, Yonsei University College of Medicine
[3]Department of internal medicine, Yonsei University College of Medicine
[4]Cardiovascular Genome Center, Yonsei University College of Medicine

Abstract

Objective: This paper aims to compare the performance of regression-based statistical approaches that were currently used or advocated to adjust a treatment effect.
Methods: The six methods used to compare their relative performance were: excluding treated individuals from data, no adjustment for treatment effect, modelling treatment as a covariate(indicator variable), non-parametric adjustment of treatment, adding a constant value to measurements for treated individuals, and censored normal regression. We applied these methods to real genetic and clinical data from Yonsei cardiovascular genome center to demonstrate a pattern of their behaviour.
Results: Two of the adjustment methods were more powerful than other methods for analysis of genetic association with serum lipid profiles. These were: no adjustment to the observed lipid profiles in treated subjects, non-parametric adjustment method based on averaging ordered residuals.
Conclusion: Non-parametric adjustment method based on averaging ordered residuals and no adjustment to the observed lipid profiles in treated subjects can effectively adjust the distorting effect of lipid-lowering drug and recover a marked loss in statistical power. Also, in genetic association studies of continuous traits that distortion arising from a treatment effect really matters, we proposed to use the appropriate methods that are more effective and straightforward to implement.

Keywords: quantitative trait, genetic association, treatment effect adjustment, regression, censored regression

* 본 논문은 보건복지가족부 보건의료바이오기술개발사업 중 특정연구센터 지원 사업(A000385)의 연구비 지원을 받아 이루어졌음.
† 교신저자 : 서울특별시 서대문구 신촌동 134 연세대학교 의과대학 의학통계학과
E-mail : biostat@yuhs.ac

1. 서 론

최근 심혈관계질환의 발생과 관련 있다고 알려진 혈압(blood pressure)이나 혈중 지질농도(serum lipid profiles)등 양적 형질의 유전적 관련성(genetic association)을 밝히고자 하는 다양한 연구들이 진행되고 있다. 그런데 연구 대상들이 혈압강하제나 지질강하제와 같은 치료 약물을 복용하는 경우, 그 영향을 배제하여 관련성을 분석하기 위해 복용대상자들을 제외하는 경우가 대부분이었다. 특히 유전적 관련성에 대한 최종 관심변수가 혈중 지질농도 그 자체가 아니라 각종 심혈관계질환의 발생 여부인 연구들에서는 연구 대상들을 수집할 때 치료 약물을 복용하는 다수의 대상들을 환자군 뿐만 아니라 대조군에 포함시켜 조사해왔다. 따라서 이런 연구들을 통해 얻어진 자료들은 치료 약물 복용 여부를 사전에 통제하여 혈중 지질농도 등의 값을 얻기가 쉽지 않은 것이 사실이다. 최근 들어, 이러한 치료 약물 등의 효과를 보정하여 양적 형질의 유전적 관련성을 분석하기 위한 통계학적 방법들이 제안되고 있는데[1], 본 연구에서는 이러한 방법들을 실제 자료에 적용하여 그 성능을 비교, 평가해보고자 한다.

2. 연구 방법

양적 형질의 유전적 관련성 연구에서 양적 형질에 영향을 미치는 요인은 크게 단일 염기 다형성(Single nucleotide polymorphisms; 이하 SNP)의 유전자형(genotype)은 다양한 다형성들의 상호 작용에 의해 양적 형질에 영향을 주는 유전적 요인(Genetic effects)과 이와 함께 연령, 성별, 흡연력 등의 환경적 요인(Environmental effects)을 나눌 수 있다. 이러한 영향력을 평가할 수 있는 방법 중 널리 쓰이는 모형으로 다음과 같은 다중 선형 회귀모형(Multiple Linear Regression)을 들 수 있다.

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad (1)$$

여기서 $\boldsymbol{y}$ 는 종속 변수인 양적 형질을 나타내고, $\mathbf{X}$ 는 양적 형질에 영향을 미치는 유전적 • 환경적 요인을 포함하는 설명변수들로 이루어진 행렬이고, $\boldsymbol{\beta}$ 는 각 요인들의 영향력을 나타내는 회귀 계수(regression coefficient)이다. 마지막 항인 $\boldsymbol{\epsilon}$ 는 추정된 회귀식으로 설명할 수 없는 부분을 나타내는 오차(error)항으로 평균이 0이고 분산이 $\sigma^2\mathbf{I}$ 인 정규 분포를 따른다. 앞으로 소개하는 여러 가지 치료 효과 보정 방법들은 모형 (1)을 기반으로 하는 것이다.

약물을 복용하거나 치료를 받은 연구 대상의 양적 형질에서 나타나는 치료 효과를 보정하는 방법을 소개하기에 앞서, 본래 양적 형질 수치(underlying quantitative trait)를 Z 라고 하고 실제 관측된 양적 형질 수치를 Y 라고 하여 치료를 받은 대상에 한하여 Y 의 보정을 통해 Z 를 추정하는 방법을 소개한다. 다양한 방법으로 보정된 Y 를 A 라고 하면 앞으로 소개할 방법들은 (1)에서 종속 변수로 A 를 이용한 모형을 기반으로 한다. 또한, 치료 여부를 나타내는 지시변수(indicator variable)로 $treat_i$ 를 사용하여 치료를 받은 연구 대상은 1, 치료를 받지 않은 대상은 0으로 나타내기로 한다.

(1) 치료 대상 제외

치료를 받은 연구 대상의 양적 형질은 편향되어 있으므로 분석 전에 제외시키고 치료를 받지 않은 대상들로만 분석하는 방법이다.

(2) 무보정

치료를 받은 연구 대상의 양적 형질에 아무런 보정을 하지 않고 그대로 분석하는 방법이다. 즉 위의 회귀 모형 (1)에서 모든 대상자의 관측된 양적 형질을 그대로 A 로 이용하면 된다.

(3) 지시변수

각 연구 대상의 치료 여부를 양적 형질에 미치는 요인으로 고려하는 모형으로, 모형 (1)에서 설명변수 행렬인 $\mathbf{X}$ 에 치료여부를 나타내는 열을 추가하는 방법이다.

(4) 비모수적 방법

이 방법은 치료를 받은 연구 대상들의 관측된 양적 형질을 토대로 계산되는 원잔차를 비모수적인 방법으로 수정하여 방법으로 Levy et al.[2]이 제안하였다. 우선 다음과 같은 영모형(null model ; $Y_i = \widehat{\beta_0} + r_i$)에 적합하여 잔차를 얻는다. 여기서 $\widehat{\beta_0}$ 는 Y_i 들의 평균을 뜻하며, i 번째 연구 대상의 원잔차인 를 구한 후 이들을 내림차순으로 정렬했을 때 k 번째에 해당하는 잔차를 r_k 로 정의한다. 여기서 k 번째 대상이 치료를 받지 않은 경우에는 보정을 하지 않고, 치료를 받은 대상인 경우는 다음과 같은 식을 통해 로 보정한다.

이는 원잔차를 내림차순으로 정렬했을 때 k 번째에 해당하는 연구 대상의 잔차와 그보다 큰 (K-1)개의 보정된 잔차들의 평균이다. 이를 치료 여부를 나타내는 지시변수($treat_k$)를 이용하여 정리하면 다음과 같다.

$$r_k^* = r_k(1 - treat_k) + treat_k\left(\frac{r_k + \sum_{j=1}^{k-1} r_j^*}{k}\right)$$

이러한 과정을 통하여 모든 연구 대상의 보정된 잔차인 를 얻은 후 내림차순으로 정렬하기 전의 순서로 다시 정렬한다. 즉 원자료의 i 번째 환자에 대하여 원잔차인 r_i 와 보정된 잔차인 r_i*를 계산하면 i 번째 연구 대상의 관측된 양적 형질은 $A_i = Y_i - r_i + r_i^*$로 보정할 수 있고 이를 종속변수로 이용하여 모형 (1)에 적합한다.

(5) 상수를 더하는 경우

치료를 받은 연구 대상에 한하여 양적 형질에 임의로 정한 상수를 더하여 분석에 사용하는 방법으로, 치료를 받은 대상들의 양적 형질은 보정하기 전의 형질의 평균이 더해지는 상수만큼 이동하지만 분포는 그대로 유지되는 장점이 있다.

(6) 중도 절단 회귀 모형

중도 절단(censoring)은 의학 분야, 특히 생존 시간을 다루는 연구에서 흔히 볼 수 있는 현상으로 일부 연구 대상의 생존 시간을 관찰하지 못하는 경우를 의미한다. 질환 연구에서 한 연구 대상의 양적 형질이 임상적으로 알려진 임의의 수치나 연구자가 미리 정해 놓은 수치보다 큰 경우 연구자의 진단 하에 치료를 받게 되며, 관측된 양적 형질은 치료에 의해 본래 양적 형질(underlying trait)보다 작아지는 효과가 발생하게 된다. 이는 관측된 양적 형질보다 본래 양적 형질은 같거나 크다고 추측하여 치료를 받은 연구 대상의 양적 형질은 우측 중도절단(right censoring)되었다고 볼 수 있다[3]. 따라서, 치료를 받은 연구 대상들은 $A_i \geq Y_i$ 로 볼 수 있고 이들의 우도(likelihood)는 다음과 같다.

$$\Pr(Y_i = y, treat_i = 1) = \Pr(A_i \geq y) = 1 - F_A(y)$$

여기서 F_A 는 A의 누적분포함수(cumulative distribution function)로 모형(1)에서 y 는 평균이 $\mathbf{X\beta}$ 이고 분산이 $\sigma^2\mathbf{I}$ 인 정규 분포를 따르므로 $F_A(y) = \Phi\left(\frac{y - X\beta}{\sigma}\right)$이다.

치료를 받지 않은 연구대상은 $A_i = Y_i$, 즉 관측된 양적 형질 Y_i 를 그대로 모형에 사용하고, 이 경우의 우도(likelihood)는 다음과 같다.

$$\Pr(Y_i = y, treat_i = 0) = \Pr(A_i = y) = f_A(y)$$

여기서 f_A 는 A의 확률밀도함수(probability density function)로 위의 F_A 에 의해 f_A 는 $\frac{1}{\sigma}\Phi\left(\frac{y - X\beta}{\sigma}\right)$ 이다.

따라서, i 번째 연구 대상의 우도를 하나의 식으로 표현하면

$$\Pr(y_i, \text{treat}_i) = [f_{A_i}(y_i)]^{1-\text{treat}_i}[1 - F_{A_i}(y_i)]^{\text{treat}_i}$$

와 같고, 총 n명의 연구 대상에 대한 우도 함수(likelihood function)을 구하면 다음과 같다.

$$L = \prod_{i=1}^{n} \Pr(y_i, \text{treat}_i) = \prod_{i=1}^{n} [f_{A_i}(y_i)]^{1-\text{treat}_i}[1 - F_{A_i}(y_i)]^{\text{treat}_i}$$

$$= \prod_{\forall i:\, \text{treat}_i=0} \frac{\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right)}{\sigma} \prod_{\forall i:\, \text{treat}_i=1} \Phi\left(-\frac{y_i - x_i'\beta}{\sigma}\right)$$

여기서, ϕ 와 Φ 는 각각 표준정규분포의 확률밀도함수와 누적분포함수이다. 대부분의 통계 패키지(package)에서는 최대우도법을 이용하여 lnL 을 최대로 해주는 β 를 뉴튼-랩슨 알고리즘(Newton-Raphson algorithm)을 통해 구한다.

앞서 언급한 방법들을 이용하여 양적 형질의 유전적 관련성을 연구하는 데에 있어 치료 효과를 보정한 결과를 비교하기 위해 심혈관계질환 유전체 연구센터(Cardiovascular Genome Center: CGC)에서 수집된 유전체 자료를 이용하였다. 이 자료의 연구 대상자는 혈연관계가 없는 총 4,214명의 개인으로 구성되어 있고, 이 중 관상동맥 조영술을 시행 받은 환자 중 적어도 한 혈관 이상에서 내경의 50% 이상 협착이 확인된 관상동맥질환군과 수축기 혈압(systolic blood pressure) 140mmHg 이상이거나 이완기 혈압(diastolic blood pressure) 90mmHg 이상, 또는 외래에서 고혈압으로 진단받은 고혈압군을 대상으로 분석하였다. 분석에서 사용할 양적 형질은 심혈관계질환의 위험 요인(risk factor)으로 알려진 혈중 지질농도 중 총 콜레스테롤(Total Cholesterol, 이하 Tchol)과 저밀도 지단백(Low-density lipoproteins, 이하 LDL) 이고, 각각의 양적 형질에 영향을 미치는 요인 중 환경적 요인은 연령, BMI, 흡연경험 유무, 음주경험유무, 운동경험 유무를 고려한다. 심혈관계질환 유전체연구센터 자료에서 알 수 있는 유전자는 총 51개에 이르기 때문에 이 들 중 관상동맥질환이나 고혈압 등 심혈관계질환의 발생 기전에 영향을 준다고 알려진 유전자를 선택하여 분석에 사용하기로 하였다. 체내에서 혈압조절 및 수분 전해질 대사에 관여하여 고혈압 발생에 중요한 역할을 하는 레닌-안지오텐신계의 구성요소 중에서는 안지오텐신 전환효소 유전자의 다형성인 14094형, G14480C, A22982G와 안지오텐시노겐 유전자의 다형성 중 G-217A, A-20C, G-6A, M235T을 택하였다[4]. 또한, 콜레스테롤 역운반과정의 주요한 역할을 하는 세포막 단백질로 알려진 ATP-binding cassette, sub-family A, member1 유전자의 다형성 R219K, 고밀도지질단백질 콜레스테롤의 대사에서 중심적인 역할을 하는 cholesterol ester transfer protein(CETP)의 다형성 중 C-239A, Taq1B, I405V, 이 밖에 관상동맥질환과 관계가 있다고 보고된 Apolipoprotein A1 유전자의 다형성 XmnI와 Apolipoprotein A5 유전자의 다형성 T-1331C, 그리고 hepatic lipase 유전자의 V95M을 유전적 요인으로 선택하였다[5,6,7,8].

치료 효과를 보정한 방법들을 적용하기 위해서는 각 연구대상의 치료 여부를 알아야 하는데, CGC자료에서는 혈압 강하제 복용 여부, 지질 강하제 복용 여부를 알 수 있었다. 본 연구에서는 지질 강하제 중 저밀도 지단백 수치의 감소에 효과적인 것으로 알려진 스타틴(statin) 계열 약물을 복용한 경우를 치료를 받은 경우로 보았다[9,10,11].

모형 (1)에 이 자료를 적합시키기 위한 회귀 모형은 다음과 같다.

$$Y_i = \beta_0 + \beta_1 age_i + \beta_2 bmi_i + \beta_3 smoke_i + \beta_4 drink_i + \beta_5 exercise_i + \beta_6 G_{1_i} + \beta_7 G_{2_i} + \epsilon_i \qquad (2)$$

설명변수 중 범주형 변수인 흡연경험유무, 음주경험유무, 운동여부, SNP는 지시변수(indicator variable)로 정의하여 모형에 포함시킨다. SNP의 유전자형은 종류가 3개인 경우 모형(2)에서 과 와 같이 2개의 지시변수로 표현할 수 있다. 예를 들어, 두 개의 대립유전자(allele) G와 A로 구성되어 있는 SNP의 경우 유전자형은 세 가지(GG, GA,

AA)이며 유전자형이 GG인 경우를 참조 범주로 정한다면 GA인 경우를 나타내는 변수를 , AA인 경우를 나타내는 변수를 로 나타낼 수 있다. 모든 분석에는 SAS ver 9.1(SAS institute, Cary, NC)를 이용하였다.

3. 결 과

관상동맥질환군과 고혈압군의 성별, 지질 강하제 복용 여부에 따른 환경적 요인 및 양적 형질의 분포는 Table 1, 2와 같다.

Table 1. Environmental factors and quantitative traits of coronary artery disease subjects

	male			female		
	Untreated(n=383)	Treated(n=383)	All(n=766)	Untreated(n=118)	Treated(n=121)	All(n=239)
variable	Mean±SD	Mean±SD	Mean±SD	Mean±SD	Mean±SD	Mean±SD
Age	58.08±9.45	55.41±8.95	56.74±9.29	60.45±9.30	60.90±8.77	60.68±9.02
BMI	24.70±2.85	25.14±2.54	24.92±2.71	24.97±3.13	25.21±2.86	25.09±2.99
	N(%)	N(%)	N(%)	N(%)	N(%)	N(%)
Smoke						
no	63(16.45)	47(12.27)	110(14.36)	109(92.37)	110(90.91)	219(91.63)
yes	320(83.55)	336(87.73)	656(85.64)	9(7.63)	11(9.09)	20(8.37)
Drink						
no	87(22.72)	107(27.94)	194(25.33)	88(74.58)	96(79.34)	184(76.99)
yes	296(77.28)	276(72.06)	572(74.67)	30(25.42)	25(20.66)	55(23.01)
Exercise						
no	173(45.29)	160(41.78)	333(43.53)	77(65.25)	71(58.68)	148(61.92)
yes	209(54.71)	223(58.22)	432(56.47)	41(34.75)	50(41.32)	91(38.08)
Traits	Mean±SD	Mean±SD	Mean±SD	Mean±SD	Mean±SD	Mean±SD
Tchol	192.07±35.77	171.07±38.00	181.56±38.35	199.58±31.36	178.59±42.27	189.00±38.64
LDL	121.58±31.56	99.31±34.52	110.32±34.90	126.38±29.07	103.39±36.85	114.73±35.11

SD: Standard deviation, Tchol: Total cholesterol, LDL: Low-density lipoproteins

Table 2. Environmental factors and quantitative traits of hypertension subjects

	male			female		
	Untreated(n=746)	Treated(n=90)	All(n=836)	Untreated(n=838)	Treated(n=109)	All(n=947)
variable	Mean±SD	Mean±SD	Mean±SD	Mean±SD	Mean±SD	Mean±SD
Age	53.03±12.07	55.17±11.35	53.26±12.00	56.24±10.66	59.54±7.96	56.62±10.43
BMI	25.20±2.99	25.48±2.75	25.23±2.97	24.89±3.10	25.34±3.09	24.94±3.10
	N(%)	N(%)	N(%)	N(%)	N(%)	N(%)
Smoke						
no	189(25.34)	22(24.44)	211(25.24)	796(95.10)	107(98.17)	903(95.45)
yes	557(74.66)	68(75.56)	625(74.76)	41(4.90)	2(1.83)	43(4.55)
Drink						
no	124(16.62)	17(18.89)	141(16.87)	606(72.32)	91(83.49)	697(73.60)
yes	622(83.38)	73(81.11)	695(83.13)	232(27.68)	18(16.51)	250(26.40)
Exercise						
no	302(40.48)	42(47.19)	344(41.20)	421(50.60)	55(50.46)	476(50.58)
yes	444(59.52)	47(52.81)	491(58.80)	411(49.40)	54(49.54)	465(49.42)
Traits	Mean±SD	Mean±SD	Mean±SD	Mean±SD	Mean±SD	Mean±SD
Tchol	198.22±37.86	184.45±40.92	196.76±38.41	210.44±40.30	193.9±44.83	208.53±41.17
LDL	122.5±34.39	104.75±36.73	120.57±35.06	132.65±35.10	115.33±40.45	130.61±36.18

SD: Standard deviation, Tchol: Total cholesterol, LDL: Low-density lipoproteins

모든 군에서 지질 강하제를 복용한 대상들의 양적 형질이 낮고 그 평균 차이는 통계학적으로 유의했으며 관상동맥질환군의 경우 약 22mmHg, 고혈압군의 경우 약 16mmHg이었다. 앞서 치료 효과를 보정하는 방법 중 다섯 번째로 설명한 상수를 더하는 방법을 적용할 때의 그 상수를 정하는 방법으로는 선행 연구를 통해 치료로 인한 양적 형질의 변화의 평균으로 보고된 값을 이용할 수 있다. 혈압과 관련된 연구를 예로 들면, Cui et al.[12,13] 는 관심 있는 양적 형질이 수축기 혈압과 이완기 혈압인 분석에서 혈압강하제의 복용에 의한 편향을 줄이기 위해 알려진 치료제 효과의 평균을 근거로 혈압 강하제를 복용한 연구 대상의 수축기 혈압에는 10mmHg을, 이완기 혈압에는 5mmHg을 더한 후 분석을 하였다. 본 논문에서는 자료를 통해 치료를 받은 대상과 그렇지 않은 대상의 양적 형질의 평균 차이를 Table 1, 2를 통해 구하여 상수를 더하는 방법을 적용할 때의 상수로 이용하였다.

모형(2)를 기반으로 유전적 요인만을 고려한 경우와 유전적 요인과 함께 환경적 요인까지 고려한 경우 치료 효과를 보정하는 6가지 방법을 적용시켰을 때 관상동맥질환군에서의 결과는 Table 3과 같고, 고혈압군에서의 결과는 Table 4와 같다. 여성 호르몬인 에스트로겐은 LDL 수용체를 통한 LDL 대사 및 콜레스테롤 대사관련 유전자들의 발현에 영향을 주는 것으로 알려져 있으므로[14] 성별에 따른 유전적인 영향을 구분하여 보기 위해 각 군의 성별에 따라 나누어 분석하였다. 표의 값들은 14개의 유전자를 이용하여 모형 (2)에 적합

했을 때 본 연구에서 고려한 14개 SNP들의 유전자형에 대한 회귀계수의 유의성 검정 결과 p-value들의 중앙값과 사분위수 범위를 나타낸다.

Table 3. Median p-value of regression coefficients of genotype in coronary artery disease subjects

	exclude	no adjustment	indicator	non-parametric	add a constant	censored regression
male						
Tchol genetic	0.5233 (0.1978, 0.7740)	0.5964 (0.3236 , 0.8103)	0.5998 (0.3359 , 0.7110)	0.4908 (0.2960 , 0.8060)	0.6038 (0.3346 , 0.7154)	0.5385 (0.2582 , 0.8154)
genetic+environmental	0.5154 (0.2459, 0.6150)	0.6723 (0.3565 , 0.8633)	0.5763 (0.3793 , 0.7270)	0.4925 (0.3307 , 0.7639)	0.5730 (0.3764 , 0.7248)	0.5886 (0.3029 , 0.7011)
LDL genetic	0.4948 (0.2532, 0.6167)	0.6328 (0.2315 , 0.6845)	0.5188 (0.2132 , 0.7737)	0.5339 (0.2749 , 0.7595)	0.5325 (0.2164 , 0.7612)	0.5893 (0.2467 , 0.7653)
genetic+environmental	0.4950 (0.2876, 0.5989)	0.6590 (0.2095 , 0.7727)	0.4097 (0.2310 , 0.7985)	0.5072 (0.2460 , 0.7137)	0.4351 (0.2302 , 0.7987)	0.4798 (0.1982 , 0.7466)
female						
Tchol genetic	0.5598 (0.4484, 0.7700)	0.5555 (0.3737 , 0.8606)	0.6066 (0.3553 , 0.8805)	0.6439 (0.3238 , 0.7607)	0.5874 (0.4322 , 0.8084)	0.6700 (0.3195 , 0.7766)
genetic+environmental	0.6372 (0.3654, 0.7816)	0.5111 (0.2937 , 0.7501)	0.5552 (0.2963 , 0.7570)	0.5666 (0.3947 , 0.7232)	0.5559 (0.3115 , 0.7663)	0.5316 (0.4124 , 0.6979)
LDL genetic	0.6898 (0.4887, 0.8867)	0.6573 (0.4272 , 0.8943)	0.7205 (0.4806 , 0.8860)	0.6221 (0.3487 , 0.8530)	0.7351 (0.5135 , 0.8737)	0.6769 (0.3763 , 0.8791)
genetic+environmental	0.6587 (0.4642, 0.8626)	0.5826 (0.4158 , 0.8899)	0.6124 (0.4330 , 0.9095)	0.5396 (0.3609 , 0.8744)	0.6666 (0.4935 , 0.8795)	0.6193 (0.3879 , 0.9085)

Tchol: Total cholesterol, LDL: Low-density lipoproteins, Numbers in parentheses are inter quartile range

Table 4. Median p-value of regression coefficients of genotype in hypertension subjects

	exclude	no adjustment	indicator	non-parametric	add a constant	censored regression
male						
Tchol						
genetic	0.3112	0.3001	0.3155	0.3891	0.3013	0.3752
	(0.1715 , 0.5532)	(0.2051 , 0.5984)	(0.2189 , 0.5463)	(0.1998 , 0.6559)	(0.2138 , 0.5230)	(0.1974 , 0.6521)
genetic+environmental	0.2895	0.2951	0.3002	0.3588	0.2937	0.3268
	(0.1416 , 0.5968)	(0.1908 , 0.5575)	(0.2109 , 0.5628)	(0.1621 , 0.6676)	(0.2035 , 0.5441)	(0.1619 , 0.6046)
LDL						
genetic	0.3681	0.3832	0.4613	0.5496	0.4407	0.4996
	(0.1691 , 0.7333)	(0.1986 , 0.6516)	(0.2156 , 0.6659)	(0.2572 , 0.7364)	(0.2158 , 0.6617)	(0.2641 , 0.7257)
genetic+environmental	0.3488	0.3528	0.3766	0.4707	0.3583	0.4061
	(0.1868 , 0.6528)	(0.1694 , 0.6441)	(0.1955 , 0.7297)	(0.2414 , 0.7470)	(0.1952 , 0.7136)	(0.2549 , 0.7328)
female						
Tchol						
genetic	0.6528	0.5464	0.5449	0.4322	0.5082	0.4418
	(0.2482 , 0.8077)	(0.1886 , 0.7264)	(0.1954 , 0.7251)	(0.2883 , 0.7231)	(0.3357 , 0.6732)	(0.3371 , 0.7016)
genetic+environmental	0.3654	0.3528	0.3553	0.3609	0.3583	0.3763
	(0.2614 , 0.8877)	(0.1700 , 0.8485)	(0.1747 , 0.8193)	(0.3081 , 0.7253)	(0.2441 , 0.7453)	(0.3624 , 0.6476)
LDL						
genetic	0.5869	0.4598	0.4965	0.4632	0.4844	0.4539
	(0.2061 , 0.7739)	(0.0774 , 0.7042)	(0.0851 , 0.7257)	(0.3047 , 0.6010)	(0.1335 , 0.7742)	(0.2497 , 0.6941)
genetic+environmental	0.4801	0.3938	0.4375	0.4516	0.5128	0.4682
	(0.1585 , 0.7990)	(0.0622 , 0.6830)	(0.0676 , 0.7344)	(0.2807 , 0.6416)	(0.1002 , 0.8093)	(0.2525 , 0.7610)

Tchol: Total cholesterol, LDL: Low-density lipoproteins, Numbers in parentheses are inter quartile range

관상동맥질환군에서 남자인 경우, 유전적 요인만을 고려하거나 환경적 요인까지 모두 고려한 경우 총 콜레스테롤에 미치는 영향을 p-value의 중앙값으로 비교하면 치료 효과를 비모수적으로 보정한 방법이 제일 작고 그 다음으로 치료 대상 제외 순이었다. 저밀도 지단백에 미치는 영향을 보면, 유전적 요인만 고려했을 때는 제외, 지시변수 사용, 상수 추가 순으로, 유전적요인과 환경적 요인을 함께 고려했을 때는 지시변수 사용, 상수 추가, 중도절단 회귀모형 순으로 나타났다. 여자의 결과는, 총 콜레스테롤에 미치는 영향을 유전적 요인만 고려하여 보면 무보정, 제외, 상수를 추가하는 방법 순으로 보이고, 환경적 요인까지 고려하면 무보정, 중도절단 회귀모형, 지시변수 사용 순이었다. 저밀도 지단백에 미치는 영향을 보면, 유전적 요인만을 고려하거나 환경적 요인까지 모두 고려한 경우 비모수적인 방법, 무보정 순으로 나타났다.

고혈압군에서 남자의 결과를 먼저 보면, 유전적 요인만을 고려하거나 환경적 요인까지 고려하여 콜레스테롤에 미치는 영향을 비교하였을 때 무보정, 상수 추가, 치료 대상을 제외 순이었다. 환경적 요인까지 고려했을 때는 제외, 상수 추가, 무보정 순이었다. 저밀도 지단백에 미치는 영향을 보면, 유전적 요인만 고려하거나 환경적 요인까지 고려한 경우 모두 제외, 무보정, 상수 추가 순으로 나타났다. 여자의 결과는, 총 콜레스테롤에 미치는 영향을 유전적 요인만 고려한 경우는 비모수, 중도절단 회귀모형, 상수 추가 순이었고 환경적 요인까지 고려한 경우는 무보정, 지시변수 사용, 상

수 추가 순으로 보였다. 저밀도 지단백에 미치는 영향을 보면, 유전적 요인만 고려했을 때는 중도 절단 회귀모형, 무보정, 비모수 순이고, 환경적 요인까지 고려하면 무보정, 지시변수 사용, 비모수 순으로 나타났다.

4. 결론 및 고찰

대부분의 연구에서 약물을 복용한 환자는 연구 대상에서 제외하는 것을 볼 수 있는데, 약물 복용을 하지 않는 대상들로만 분석하면, 이는 전체 환자 중 약물을 복용한 환자의 비율이나 약물 복용 여부에 따른 형질의 차이 등으로 인해 자료의 손실을 가져오고 결국은 연구 결과에 왜곡을 가져오게 된다. 또한, 약물을 복용하는 환자들은 대부분 본래 형질은 관측된 형질보다 클 것으로 추측할 수 있는데, 이들을 제외한다면 이러한 형질의 정보 측면에서도 손실이 발생한다. 따라서 이러한 약물 복용으로 인한 효과를 보정하여 연구 결과에 연구 대상의 본래 특성을 반영하고자 하여야 한다.

치료 여부를 지시 변수로 나타내어 이를 양적 형질에 영향을 주는 환경적 요인으로 모형을 설정하는 방법은 잘못된 것이다. 우선, 치료나 약물 복용은 본래 지질농도나 혈압 등 양적 형질이 높은 환자에게 처방되기 때문에 이는 양적 형질에 의한 결과물일 뿐, 양적 형질이 종속 변수인 회귀모형에서 설명 변수에 포함되어 양적 형질에 영향을 미치는 요인으로 볼 수는 없다. 만약 양적 형질에 아무런 보정 없이 치료 여부를 설명 변수로 삽입한다면 치료를 받은 대상의 양적 형질은 이미 치료로 인해 본래의 양적 형질보다 작아졌으므로 그에 대응되는 회귀계수가 음수일 가능성이 높다. 회귀 식의 양변에서 치료여부를 나타내는 지시변수 항을 빼주면 이는 설명한 상수를 더하는 방법에서 상수를 자료를 통해 얻고 그 값이 음수인 경우와 동일한 것을 알 수 있다. 실제 분석결과에서도 무보정보다 보정력이 떨어짐을 볼 수 있었으므로 분석하기에는 쉬운 방법이지만 통계학적으로 좋은 방법이라고 보기 어렵다.

상수를 더하는 방법을 이용하여 치료 효과를 보정하고자 할 때 선행 연구 등을 통한 치료 효과에 대한 정보가 없는 경우, 치료를 받은 대상의 형질은 관측된 형질을 기준으로 우측 중도 절단되었다고 가정하고 회귀분석을 할 수 있어 유용한 방법이 중도 절단 회귀모형이다. 하지만 이 모형은 중도 절단여부, 즉 치료 여부가 무정보적(non-informative)이어야 한다는 가정이 전제되어야 한다. 만약 어떤 약물에 대한 임상 연구에서 연구대상이 약물의 부작용을 알고 추적 조사를 중간에 포기한다면 이는 정보적 중도절단(informative censoring)에 해당한다. 또한, 우측 중도 절단이 가정이므로 약물을 복용한 대상의 본래 형질은 관측된 형질보다 최소한 같거나 크다는 가정이 있어야 한다.

비모수적으로 잔차를 수정하여 형질을 보정하는 방법은 치료를 받은 대상에 한해서만 관측된 형질보다 큰 값을 가지는 부분을 이용한다는 점에서 중도절단 회귀모형에서 구하는 우도 함수와 그 원리가 비슷하다. 또한, 중도절단 회귀모형은 오차의 정규성을 가정하지만 비모수적으로 잔차를 수정하는 방법을 적용시킬 때에는 정규분포를 따른다는 가정이 필요 없다.

본 연구에서 SNP의 유전형은 혼합형 유전 모형(codominant model)로 가정하여 3개의 유전형의 평균이 모두 다르다고 보고 이 중 2개의 유전형에 대한 지시변수를 만들어서 양적 형질과의 유전적 관련성을 보았는데, 우성 모형(dominant model)이나 열성 모형(recessive model)으로 가정하여 특정 대립유전자의 양적 형질과의 유전적 관련성을 여러 보정 방법을 적용시켜서 비교해 볼 수 있을 것이다. 또한 약물로 인해 변화한 양적 형질 수치를 비율로 계산하여 이를 역산하는 방법으로 보정하는 방법도 고려할 수 있다. 예를 들어, 약물로 인해 양적 형질이 본래 연구대상의 양적 형질보다 $(100 \times a)\%$ 감소한다면, 약물을 복용한 대상에 한하여 관측된 양적 형질에 $\frac{1}{1-a}$를 곱하여 보정

해주는 방법을 고려할 수 있을 것이다.

본 논문에서는 제안된 방법들을 이용하여 실제 자료에 적용해보고 모형에서의 회귀계수 유의성 검정에 대한 p-value로 모형들을 비교하였다. 하지만 모든 경우에서 p-value가 상당히 큰 결과를 보여 p-value만으로 제안된 방법들을 비교 및 평가하는 데에 어려움이 있었다. 따라서 회귀 분석에서의 모형의 적합성을 나타내는 AIC(Akaike's information criterion)나 R^2, 즉 결정계수를 이용하여 모형들을 비교해볼 수 있을 것이다.

참고문헌

[1] Tobin MD, Sheehan NA, Scurrah KJ, Burton PR. Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure. Stat Med. 2005;24:2911-2935.

[2] Levy D, DeStefano AL, Larson MG, et al. Evidence for a gene influencing blood pressure on chromosome 17. Genome scan linkage results for longitudinal blood pressure phenotypes in subjects from the Framingham Heart Study. Hypertension 2000; 36(4):477 - 483.

[3] Cook NR. An imputation method for non-ignorable missing data in studies of blood pressure. Stat Med. 1997; 16(23):2713 - 2728

[4] Kim JM, Shin DJ, Yoon BJ, et al. Association between I/D, G14480C, A22982G Polymorphisms of Angiotensin I-Converting Enzyme Gene and Essential Hypertension in the Korean Population. Korean Circ J 2004;34:1137-1147

[5] Ko YG, Cho EY, Park HY, et al. Association of R219K Polymorphism in the ABCA1 Gene with Plasma Lipid Levels and Coronary Artery Disease in Koreans. Korean Circulation J 2003;33(1):44-51

[6] Moon JY, Cho EY, Kim WH, et al. The Impact of Apolipoprotein A-I Polymorphisms on the Lipid Profiles in Middle Aged Healthy Men and Women. Korean Circ J 2004;34:1158-1166

[7] Song KJ, Lim KS, Cho JN, et al. Linkage Disequilibrium Analysis of Quantitative Trait Locus Associated with Lipid Profiles. Korean Circ J 2006;36:688-694

[8] Cho EY, Bae SJ, Cho HK, et al. Association of Cholesteryl Ester Transfer Protein Gene Polymorphism with Serum Lipid Concentration and Coronary Artery Disease in Korean Men. Korean Circ J 2004;34:565-573

[9] Kim CH, Kim KI, Kim JH, et al. Prospective Study to Evaluate the Efficacy and Safety of Pitavastatin in Patients with Risk Factor of Cardiovascular Disease(PEACE Study). Korean Circ J 2007;37:16-21

[10] Chun KJ, Chung NS, Ahn SK, et al. Efficacy and Safety of Atorvastatin in Patients with Elevated LDL-Cholesterolemia. Korean Circ J 1999;29:1309-1316

[11] Jun JE. Cholesterol Lowering Therapy in Coronary Artery Disease - With Particular Reference to Statins -. Korean Circ J 2001;31:849-856

[12] Cui J, Hopper JL, Harrap SB. Genes and family environment explain correlations between blood pressure and body mass index. Hypertension 2002; 40(1):7 - 12.

[13] Cui JS, Hopper JL, Harrap SB. Antihypertensive treatments obscure familial contributions to blood pressure variation. Hypertension 2003; 41(2):207 - 210.

[14] Kim JW, Song JH, Kim YM. Regulation of cholesterol homeostatis-related gene expression by high fat diet and estrogen. J lipid vasc dis. 2000; 10(2):177-187.