creative commons
COMMONS DEED

저작자표시 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.
- 이차적 저작물을 작성할 수 있습니다.
- 이 저작물을 영리 목적으로 이용할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

이학석사 학위논문

High-order gene-gene interaction analysis in Genome Wide Association Studies

전장유전체연구에서의 고차원 유전자-유전자 간의 교호작용 분석

2013년 2 월

서울대학교 대학원

통계학과

김 용 강

High-order gene-gene interaction analysis in Genome Wide Association Studies

지도 교수 박 태 성

이 논문을 이학석사 학위논문으로 제출함
2013 년 2 월

서울대학교 대학원
통계학과
김 용 강

김용강의 이학석사 학위논문을 인준함
2013 년 2 월

위 원 장 박 병 욱 (인)

부위원장 박 태 성 (인)

위 원 Paik Myunghee Cho (인)

Abstract

Genome-wide association studies already have found hundreds of associations between genetic variants and complex human diseases and traits. Most GWA-studies are concentrated on single variants size. It was shown that most variants found by GWA-studies could explain a small part of human diseases heritability. Consequently, many researchers have studied gene-gene or gene-environment interactions. In 2001, Ritchie et al. proposed MDR method for the determination of gene-by-gene and gene-by-environment interactions. This method has a benefit of fitting and interpreting effects of gene-gene interactions. However, this method can only be applicable to case- control data and cannot adjust for other environmental variables. To overcome these disadvantages, Xinag-Yang et al. proposed generalized MDR method in 2007. Despite of its benefits, this method also has a data sensitive problem. In other words, performance of GMDR method is affected by erroneous samples Erroneous samples means the ones that are diverged from the general tendencies of other samples. In this reason, we first consider the effects of erroneous samples in GMDR analysis and propose two methods for reducing the effects caused by erroneous samples. Methods to reduce effects of erroneous samples which we propose are L-estimator GMDR and M-

estimator. L-estimator and M-estimator are statistical methods deriving for robust estimation. In this study, to adjust concepts of L-estimator and M-estimator to GMDR method has advantages in consistency of choices. As a result we reveal that L-estimator GMDR and M-estimator has a benefit to robustness by simulation and real data analysis.

주요어 : Gene-Gene interaction, MDR, GMDR, robustness, L-estimator, M-estimator
학 번 : 2011-20241

Contents

List of Tables

List of Figures

Introduction

With the development of generating and handling genomic data, a genome-wide association (GWA) study has become a common approach for testing association between a single nucleotide polymorphism (SNP) and a complex disease of interest.[1] There have been many successful results from GWAS. However, SNPs which were found by GWAS are shown to explain only a small fraction of disease etiology due to ignoring relatedness between complex diseases and multiple genes and/or their interactions. In this reason, Analysis of gene-gene and/or gene-environment interactions has been emphasized as a new alternative for understanding the etiology of common complex disease. However, Gene-gene interaction is hard to detect and characterize by using traditional parametric statistical methods with follow reasons.[2] First, the sparseness of the data is occurred in ultra-high dimensions. For overcoming sparseness of the data, parametric statistical methods for finding gene-gene interactions need to exponentially larger sample sizes. Second, detecting gene-gene interactions using traditional procedures leads to an increase in type II errors and a decrease in power. As a result, detecting interactions among variables is a well-known challenge in statistics and data mining (Freitas, 2001).[3] For detecting

gene-gene interaction, Ritchie et al. proposed Multifactor Dimensionality Reduction(MDR) method.[4] MDR method isn't affected by sparseness problem because its procedure allows multilocus genotype combinations that have very few or no data points. However, original MDR method can be only used in association studies of discrete traits. To adapt MDR method for using in continuous traits, Xiang-Yang et al. proposed Generalized Multifactor Dimensionality Reduction(GMDR) method.[5] GMDR method has advantage to sparseness problem, therefore GMDR method can be used for detecting gene-gene interaction in association studies of continuous traits.

For developing original MDR and GMDR method, there are a lot of versions of MDR developed by many researchers. Chung et al. proposed odds ratio based MDR(OR-MDR) for categorizing risk level precisely.[6] Calle et al. proposed model based MDR(MB-MDR) for parametric extension of MDR method.[7] Gui et al proposed survival MDR(Surv-MDR) for survival data analysis.[8] Lee et al proposed Cox-MDR for semi-parametric survival data analysis.[9] Oh et al. proposed gene-based MDR for investigating gene-gene interaction by interaction of multiple locus set.[1]

These methods have brought many remarkable successes in gene-gene interaction analysis. However, there are no researches about the effects of erroneous samples to results of

MDR methods. Erroneous samples means which diverge from the usual tendencies of other samples which expected to be similar with them. Outliers of regression models are representative example. Unfortunately, there is regression step for performing GMDR. In this reason, we first show the effects of erroneous samples and propose two methods for reducing effects caused by erroneous samples.

Materials and Methods

KARE data

The Korea Association Resource (KARE) is a project for gathering large-scale GWA analyses in the Gyeonggi Province of South Korea.[10] There are two community cohorts that participated in KARE project: Ansung and Ansan cohorts. The number of people in Ansung cohort is 5018 and the number of people in Ansan cohort is 5020. Age is distributed from 40 to 69, and more than 260 traits have been extensively examined through epidemiological surveys, physical examinations and laboratory tests. In this paper, we use the Homeostasis Model Assessment of Insulin Resistance(HOMA-IR) levels which is widely used to estimate insulin resistance. Since HOMA-IR has a nonnegative skewness distribution, the gamma distribution is commonly assumed.[11] DNA samples were genotyped on the Affymetrix Genome-Wide Human SNP array 5.0 which can genotype 500,568 SNPs. We used quality control processes which mentioned in first paper of KARE project.[10, 12] We deleted samples whose BMI or HOMA-IR scores are missing. After filtering the samples with the missing phenotypes, a total of 8,577 individuals and 327,872 SNPs were remained in the analyses.

MDR method

MDR method has several steps to evaluate genetic effects. The first step is selecting genetic factors which are evaluated by their effects on phenotype. If we want nth order gene-gene interactions, we select n SNPs and make 3^n contingency table. In each cell of contingency table shows the numbers of cases and controls which have same genetic factor combination. The second step calculates ratio of case to control in each cells, and determine "high risk" or "low risk" cell by their ratio of case and control. If the ratio of case to control in cell exceeds some threshold (commonly used 1), the cell is labeled as "high risk". If not, the cell is labeled as "low risk". Through such process, each cell can be summarized as "high risk" or "low risk". In the third step, the selected SNP sets' performances of determination of case or control are measured. If total samples' ratio of case to control is almost 1, accuracy function is used to measure for determining performance. Accuracy function measures the performance by ratio of number of true determination samples to number of total determination samples. However, if our sample is not balanced Chawla et al.[13] concluded accuracy function is biased in terms of measuring determination performance. Such problem led Velez et al. to

propose a balanced accuracy function.[14] Balanced accuracy function is 1/2(sensitivity+specificity), where sensitivity is ratio of number of true determination case samples to number of total case samples, and specificity is ratio of number of true determination control samples to number of total control samples. In the 4th step, we select best performance combination of SNPs by adjusting above steps to different SNP sets and comparing SNP sets by accuracy or balanced accuracy. However, if we perform these steps as given, there is over fitting problem which means our selected SNP sets' performance measured higher than the actual performance. To overcome this problem, we use cross-validation method, before we perform 1st to 4th steps and we select best combination of SNP sets in each cross-validation sets. In the 5th step, we use cross-validation consistency(CVC) measure for finding best SNP sets across cross-validated samples

GMDR method

Although there are a lot of research achievements by using MDR method, MDR has several shortcomings. First of all, MDR method can only adjust for classification of case: control data, because MDR method classifies each cell by ratio of case to control. Second, the MDR method cannot adjust other environment or genetic factors. If environment factors or genetic factors which affect to phenotype and which are affected by sampling exist, they might cause biased result to MDR method. For overcoming these problems, Xiang-Yang et al. proposed Generalized Multifactor Dimensionality Reduction(GMDR) method.[5] Steps of GMDR method are very similar to original MDR method. However, in GMDR method, the score is used instead of the counts of individuals. Xiang-Yang et al. proposed residual based score in their first GMDR paper. Let y_i denote the phenotype of individual i, in this paper, phenotype is continuous although it should not be continuous type. Let $E(y_i) = \mu_i$. In general, we want to find significant vector β, which is the model below.

$$l(\mu_i) = \alpha + x_i^T\beta + z_i^T\gamma \quad (1)$$

Where $l(\mu_i)$ is link function, α is the intercept, and x_i is the vector of possible expression of genotype combination of interest. z_i is the vector of environment factors which we want to adjust in model, and β and γ are coefficient vectors. Since our research concentrate on continuous phenotype, our link

function is identity and we supposed that y_i is distributed as independent normal distribution with equal variance. As figure1, we fit the models with environmental variables and responses with all samples, it means we assumed model follows null hypothesis such that β equals 0 vectors. It means if this fitting can explain little bit of samples, there are some gene effects to explain phenotype. In conclusion, we need to find the gene-gene interaction which can explain residual part a whole. In this reason, we denote score to each individual by residual. Xiang-Yang Lou et al. proposed Pearson residual because it is derived from score statistic. In 2^{nd} step, similarly to MDR method, we fill the cells by individuals who have these allele combinations. And we summed up the score as in figure3 and if this score is larger than threshold(commonly used 0), we denote this cell "high risk", and smaller than threshold, then we denote this cell "low risk". In figure3, "aabb" cells' sum of the score is higher than 0, we denote "aabb" cell "high risk". Next steps are same as MDR method. GMDR method has a benefit comparing to MDR method, because GMDR can be used to continuous response variable and model which adjusting environmental variables. However there are some problems using GMDR method. In MDR method, individuals in equal cell have an equal weight. In GMDR method, however, individuals in equal cell have different weights. From this property, GMDR method is relatively more sensitive in existing outlier individuals.

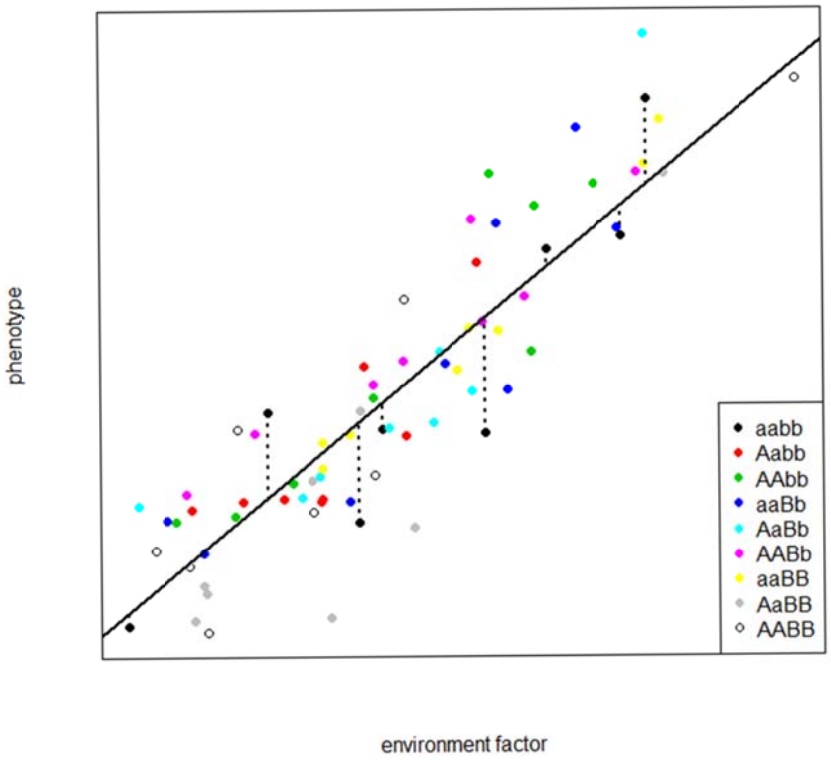


Fig 1. Example of step1 of GMDR method.
First step for GMDR method is fitting regression model with fitting phenotype with environment factor. Black dots are residuals of individuals whose combination of genotype is "aabb".

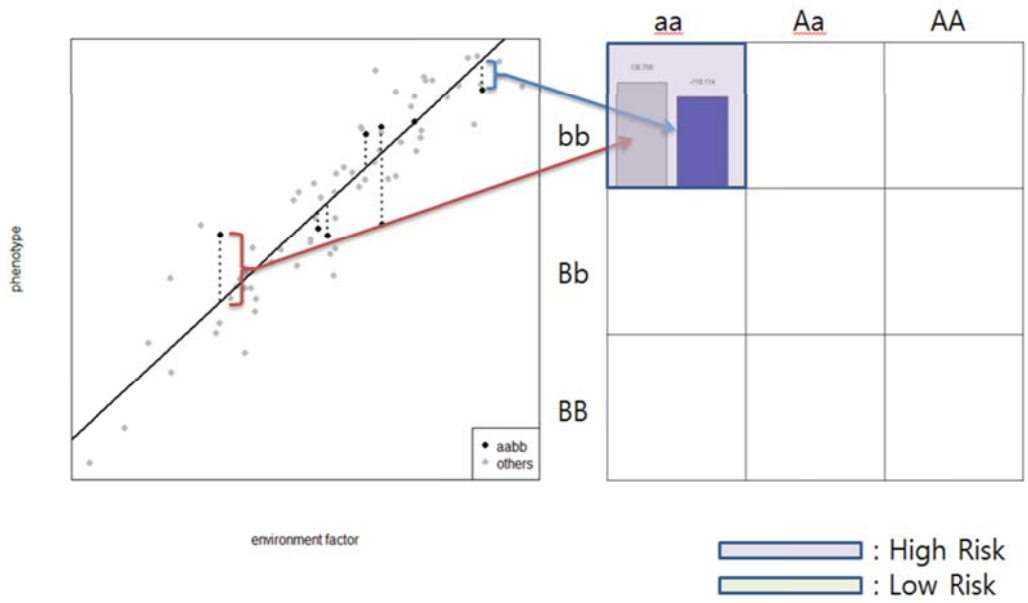


Fig 2. Description of 2nd step of GMDR method

In 2nd step of GMDR method, we used sum of the pearson residuals to each combination of genotypes. After sum up the residuals, comparing with threshold and determine high risk or low risk cells.

Robust GMDR

In this paper, for overcoming sensitive problem of GMDR method, we propose two types of scores. One of the residual score is L−estimator type score. L−estimator is an estimator that equals a linear combination of order statistics which we want to measure. One of the famous L−estimator is least trimmed square(LTS).[15] LTS regression is formulated as follows.

$$argmin_{\beta} \sum_{i=1}^{k} |r_{(i)}(\beta)|^2 \qquad (2)$$

$$where\ r_i(\beta) = \left(y_i - f(x_i, \beta)\right),\ i = 1, \ldots, n$$

In above formula, k is the number of the samples which we actually used in regression and n is overall samples. $r_{(i)}(\beta)$ means ith residual for fixed β which ordered by absolute value of residuals. Bickel and Lehmann said that trimmed expectations are the only ones which are both robust and whose estimators have guaranteed high efficiency. Such property, LTS has good performance comparing with other least square estimator. As LTS regression, when we define trimmed residual as follows and use trimmed studentized residual as score for GMDR. Since our aim to design this score is reduction of outlier's effects, we think that using studentized residual is more adequate than using pearson residual, because our model is based on linear regression. Our trimmed

studentized residual is shown as follows.

$$\dot{r}_i(\beta) = \begin{cases} \frac{y_i - f(x_i,\beta)}{\sqrt{var(\widehat{y_i - f}(x_i,\beta))}} & if \left| \frac{y_i - f(x_i,\beta)}{\sqrt{var(\widehat{y_i - f}(x_i,\beta)}} \right| \le k \\ 0 & if \left| \frac{y_i - f(x_i,\beta)}{var(\widehat{y_i - f}(x_i,\beta)} \right| > k \end{cases} \quad (3)$$

In this score formula, k is the threshold to determine whether using ith sample or not. By tuning k, we can determine how strong trim outliers. We define L-estimator GMDR which uses trimmed studentized residual as score.

The second score is M-estimator type score, or we call this score threshold residual score. M-estimator is an estimator, which is obtained as the minima of sums of functions of the data. Least-squares estimators and many maximum-likelihood estimators are M-estimators. Concept of M-estimator is derived from robust regression by Huber.[16] Tukey proposed Tukey's biweight function which is type of M-estimator for robust regression. Solving M-estimator by Tukey's biweight function is given defined as follows.

$$\rho(\mathrm{r_i}(\beta)) = \begin{cases} \frac{1}{6}[1 - (1 - \mathrm{r_i}(\beta)^2)^3] & if\ |\mathrm{r_i}(\beta)| \le 1 \\ \frac{1}{6} & if\ |\mathrm{r_i}(\beta)| > 1 \end{cases}$$

$$argmin_\beta \sum_{i=1}^{n} \rho(r_i\ \ (\beta)) \quad (4)$$

$$where\ r_i(\beta) = \big(y_i - f(x_i,\beta)\big),\ i = 1, \dots, n$$

Since solving M-estimator is based on differentiation of ρ function, biweight function's shape was made for being able to differentiation. To use idea of biweight function, we define the threshold residual score given by

$$\dot{r}_i(\beta) = \begin{cases} \frac{y_i - f(x_i,\beta)}{\sqrt{var(\widehat{y_i - f}(x_i,\beta)}} & if \left| \frac{y_i - f(x_i,\beta)}{\sqrt{var(\widehat{y_i - f}(x_i,\beta)}} \right| \leq k \\ k \cdot sign\left(\frac{y_i - f(x_i,\beta)}{\sqrt{var(\widehat{y_i - f}(x_i,\beta)}} \right) & if \left| \frac{y_i - f(x_i,\beta)}{\sqrt{var(\widehat{y_i - f}(x_i,\beta)}} \right| > k \end{cases} \quad (5)$$

In this formula, we use the threshold k to shrink residuals exceeding k. Since we do not need to differentiate function, we modify biweight function simply reducing weight of extreme loss.

Results

Simple case simulation

Before extensive simulation, we make simple example which demonstrate outliers' effect in score-based GMDR.

We assumed model as follows.

$$y = envir_1 + 0.2SNP_1 + 0.2SNP_2 + 0.18SNP_1SNP_2 + \epsilon \quad (6)$$

$$envir_1 \sim N(0,1),\ SNP_1, \ldots, SNP_{100} \sim Bin(2,0.3),\ \epsilon \sim N(0,1)$$

We use additive coding method to generate samples. We generate 2990 samples by model (5) and we generate 10 samples by follow model (6).

$$y = envir_1 + 0.2SNP_1 + 0.2SNP_2 + 0.18SNP_1SNP_2 + 2 \quad (7)$$

$$envir_1 \sim N(0,1),\ SNP_1, \ldots, SNP_{100} \sim Bin(2,0.3)$$

By this outlier generation formula, we can generate some outliers in the sample. As a result, Distribution of samples is mixed, therefore, we assign data with all samples "mixed data", and we assign data without samples which are generated by distribution (6)"pure data". We generate 998 SNPs data($SNP_3, \ldots, SNP_{1000}$) which is not relevant to phenotype for testing score based GMDR's power of detecting true effect SNP sets. For simple, we assumed all of the SNPs are mutually independent and have same minor allele frequencies(MAF), 0.3. We performed score-based GMDR method with 10 fold cross-validation.

In summary, we generate 2990 normal samples and 10 outlier

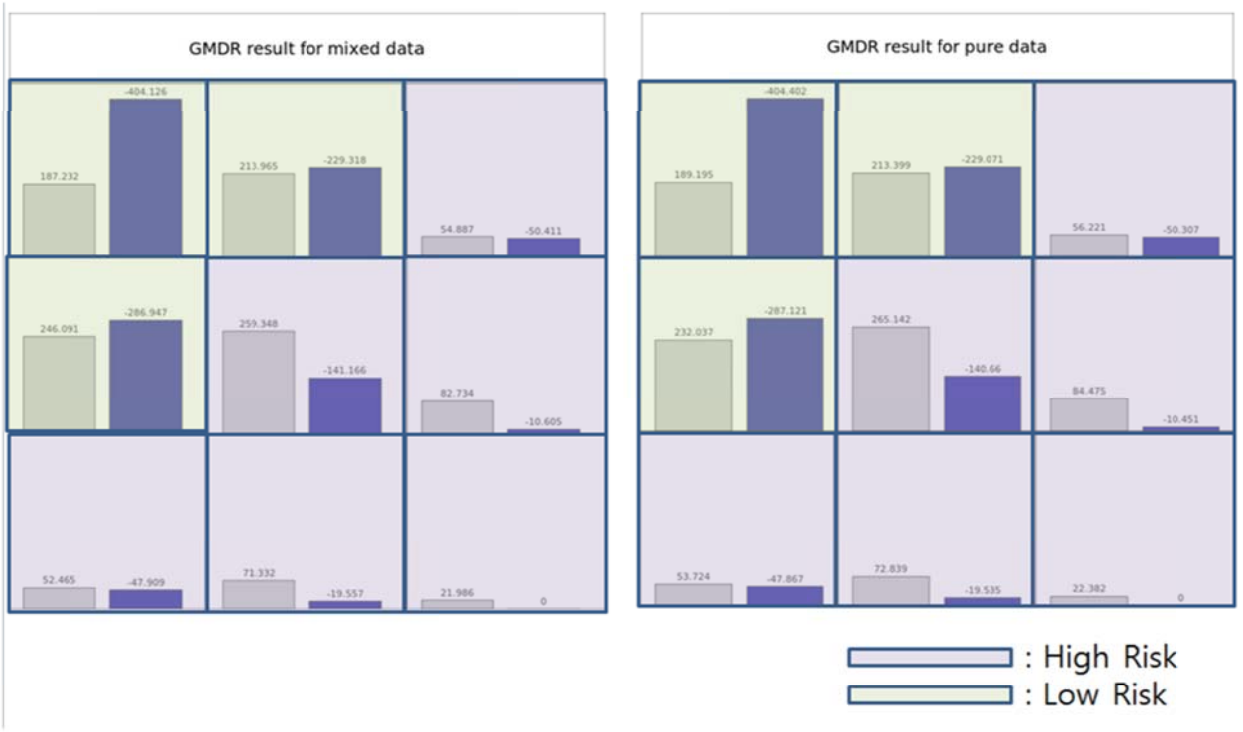


Fig 3. Result of toy example.

This 3X3 contingency tables show that there are no significant changes between table of using mixed samples and table of using only pure samples

samples. There are two effect SNPs and 998 independent SNPs with phenotype. Since toy example is just simple example for arguing the problem, we perform above situation once.

Figure 3 shows two 3x3 contingency tables which are the distribution of residual sum of each cell. These two tables were made by two causal SNPs. The left table shows result of using all samples(mixed samples) to fitting GMDR. In contrast, right table shows the result of using 2990 samples without generated from (6) distribution(pure samples) to fitting GMDR. There are no differences between two tables significantly.

Table1. Results of toy example

	Best_Combi	CVC	Aver_Train_BA	Aver_Test_BA
Mixed sample	**(SNP1 SNP979)**	**3**	**0.5789**	**0.5621**
Pure sample	**(SNP1 SNP2)**	**6**	**0.5798**	**0.5759**

However, in table1, there is quite different power between two tables. In table 1, GMDR on mixed data fails to detect true SNP sets(SNP_1, SNP_2) which affect to phenotype and has low cross-validation consistency(CVC) score. In contrast, GMDR by pure data successfully detect true SNP sets and has high CVC score. It means that outliers lead GMDR to lose detection ability. This fact also can be found in difference between training and test balanced accuracy(BA). Training BA shows interpretation power of model, therefore, in toy example, GMDR by mixed sample and GMDR by pure sample have almost same interpretation power. Test BA shows prediction power of model,

therefore, in toy example, GMDR by mixed samples has lower power than GMDR by pure samples. In this reason, we can assume that outliers caused loss of prediction power of GMDR and it makes loss of consistency in decision of GMDR method.

General case simulation

First, we generated general samples by follow model. Distribution of general samples are almost same as previous simple case's settings

$$y = envir_1 + \beta_1 SNP_1 + \beta_2 SNP_2 + \beta_{1\times 2} SNP_1 SNP_2 + \epsilon \quad (8)$$

$$envir_1 \sim N(0,1),\ SNP_1, \ldots, SNP_{1000} \sim Bin(2,0.3),\ \epsilon \sim N(0,1)$$

We generate latent outlier samples by the following model

$$y \sim N(0,400) \quad (9)$$

$$envir_1 \sim N(0,1),\ SNP_1, \ldots, SNP_{1000} \sim Bin(2,0.3),\ \epsilon \sim N(0,1)$$

Of course, some of the latent outlier samples may not behave outlier. However we simulated by different n, m, betas. We generated total 3000 samples per each iteration, and compare performance of original score based GMDR, L-estimator based GMDR, and M-estimator based GMDR. We repeated 1000 times for calculating empirical power and average CVC. Empirical power estimated as follows.

$$\frac{1}{1000}\sum_{i=1}^{1000} I(set\ \{SNP_1, SNP_2\}\ is\ selected\ in\ ith\ iteration) \quad (10)$$

We calculated average BA and CVC by 10-fold cross validation.

Our simulation's detail settings are in table 2. The first to third columns of table2 denote the effect sizes of SNP1, SNP2 and interaction in pure samples respectively.

Table 2 Models of simulation

	SNP1	SNP2	interaction	mixed proportion
Design1	**0.2**	**0.2**	**0.18**	**0.067**
Design2	**0.2**	**0.2**	**0.18**	**0.033**
Design3	**0.4**	**0.4**	**-0.3**	**0.067**
Design4	**0.4**	**0.4**	**-0.3**	**0.033**
Design5	**0.25**	**0.25**	**0.15**	**0.067**
Design6	**0.25**	**0.25**	**0.15**	**0.033**
Design7	**0.2**	**0.2**	**0.18**	**0**

The fourth column of table2 indicates mixed proportion. Mixed proportion means number of samples generated by (8) model per number of all samples. If mixed proportion is larger, simulation design generates more outlier. Rows of column denote each design of simulation. Design1, 2 and 7 have the same conditions without mixed proportion. Relationships between settings of design 3 and 4 or between settings of design 5 and 6 are the same as relationship between design 1, 2 and 7.

Table3 Results of simulation I

Empirical power	score-based GMDR	L-estimator gmdr	M-estimator gmdr
Design1	**0.299**	**0.542**	**0.523**
Design2	**0.467**	**0.705**	**0.718**
Design3	**0.289**	**0.624**	**0.566**
Design4	**0.393**	**0.691**	**0.703**
Design5	**0.363**	**0.589**	**0.58**
Design6	**0.511**	**0.745**	**0.742**
Design7	**0.847**	**0.845**	**0.848**

Table 4 Results of simulation II

Average CVC	score-based GMDR	L-estimator GMDR	M-estimator GMDR
Design1	**9.0334**	**9.2601**	**9.1816**
Design2	**9.3041**	**9.4170**	**9.3719**
Design3	**8.3356**	**8.9054**	**8.6025**
Design4	**8.7048**	**9.1360**	**8.8407**
Design5	**9.0110**	**9.3447**	**9.2724**
Design6	**9.3268**	**9.4107**	**9.4596**
Design7	**9.4911**	**9.4391**	**9.4764**

Table 3 and 4 are the result for several simulation situations. Each column of table3 indicates empirical power of each GMDR method and each column of table4 indicates average CVC of each GMDR method. By comparing designs, we can find out if we have the same effect size of SNPs, all of the results from little mixed proportion of samples make more power. In design 1 to 6, score-based GMDR has the worst power to detect

correct SNP set clearly. Only in design7, score-based GMDR reached similar power other methods. These described above tendencies are remained in average CVC(table 4). L-estimator based GMDR and M-estimator based GMDR have similar power. However there is tendency that L-estimator based has more power in designs which have high mixed proportion than M-estimator based GMDR. In contrast, M-estimator based GMDR has more power which have low mixed proportion than L-estimator based GMDR.

Real data analysis

We applied our proposed method to KARE data with HOMA-IR phenotype. Figure4 shows histogram and summary statistics of HOMA-IR. Left histogram and summary statistics in figure4 show distribution of HOMA-IR without log-transformation. Many researchers performed log-transform before regression analysis because of skewed distribution of HOMA-IR.[11] Thus we also performed log-transform and results are in the right side of figure4. In spite of performing log-transform, distribution of HOMA-IR remained skewed. Figure5 is the QQ-plot of HOMA-IR for checking normality, left side is result of without log-transform and right side is result of with log-transform. The distribuion of with log-transform seems to be more adequate to normality assumption than result of without log-transform. Before performing GMDR, we performed single SNP analysis for reducing computation burden of GMDR. Single SNP analysis with regression presented below is implemented for all SNPs.

$$\mathrm{HOMA-IR_i} = \alpha_0 + \gamma_1 SEX_i + \gamma_2 AGE_i + \gamma_3 AREA_i + \gamma_4 BMI_i + \beta_j SNP_{ij} + \varepsilon_i \quad (11)$$

$$\log(\mathrm{HOMA-IR_i}) = \alpha_0 + \gamma_1 SEX_i + \gamma_2 AGE_i + \gamma_3 AREA_i + \gamma_4 BMI_i + \beta_j SNP_{ij} + \varepsilon_i \quad (12)$$

Note that $\mathrm{i} = 1, \ldots, 8577$ and 8577 is the number of individuals; $\mathrm{j} = 1, \ldots, 327{,}872$ and 327,872 is the number of all SNPs. All

SNPs are ranked in ascending order of the p-value. We selected top 1000 ranked SNPs for performing GMDR.

Note we perform regression analysis with HOMA-IR. We used sex, age, area and BMI as environment covariate. The regression nidels are as follows.

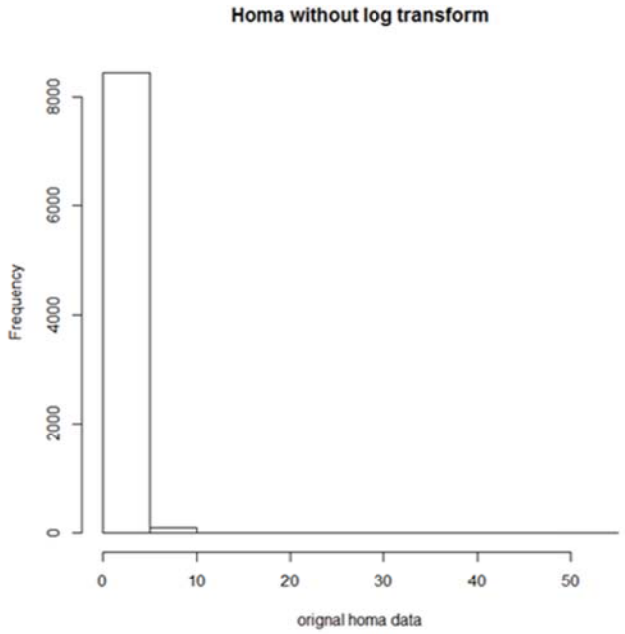
$$\mathrm{HOMA-IR_i} = \alpha_0 + \gamma_1 SEX_i + \gamma_2 AGE_i + \gamma_3 AREA_i + \gamma_4 BMI_i + \varepsilon_i \quad (13)$$

$$\log(\mathrm{HOMA-IR_i}) = \alpha'_0 + \gamma'_1 SEX_i + \gamma'_2 AGE_i + \gamma'_3 AREA_i + \gamma'_4 BMI_i + \varepsilon'_i \quad (14)$$

If we used HOMA-IR values as a response, we calculate scores based on formula (11). If we used log(HOMA-IR) values as response, we calculate scores based on formula (12). We assign k by 3, in L-estimator, and we deleted below 2% samples which have residual scores exceed threshold. We performed score-based GMDR, L-estimator GMDR and M-estimator GMDR to log-transformed HOMA-IR and log(HOMA-IR) respectively.

Min.	1st Qu	Mean	3rd Qu.	Max.
0.017	1.043	1.669	2.044	53.91

Min.	1st Qu	Mean	3rd Qu	Max.
-4.102	0.042	0.316	0.715	3.987



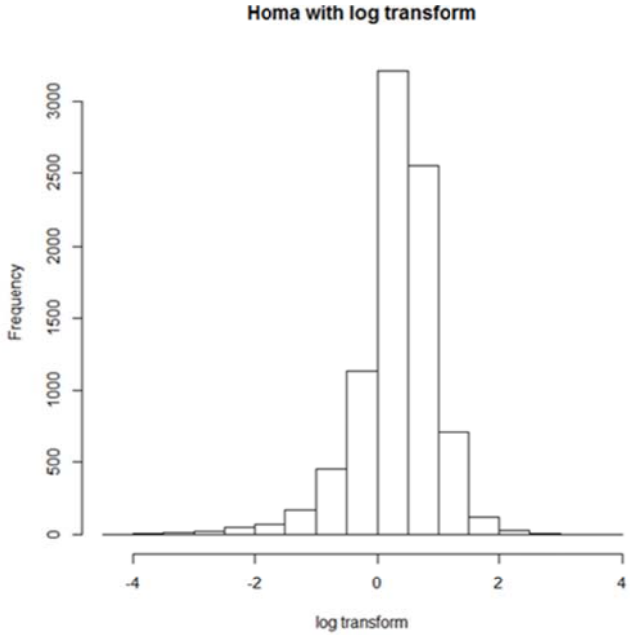


Fig4. Distribution of HOMA-IR

Before log transformation to HOMA-IR, HOMA-IR has much skewed distribution. In contrasts, after log transformation, HOMA-IR has almost symmatric distribution.

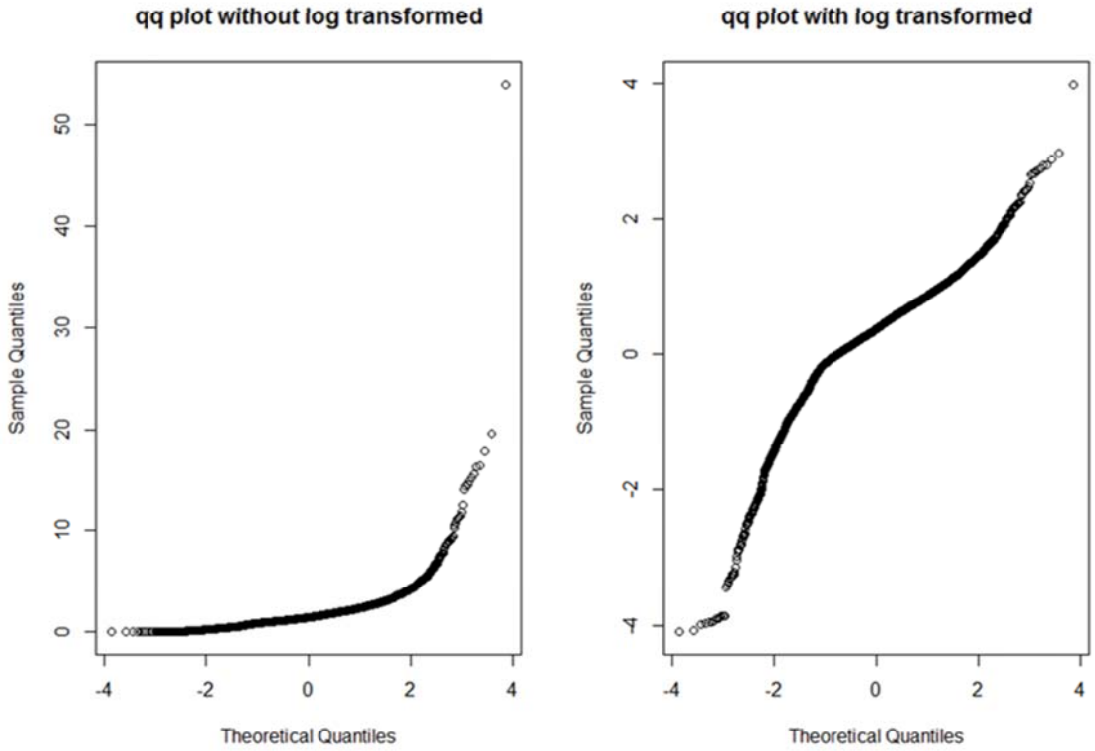


Fig 5. QQ-plot for HOMA-IR data

This figure show QQ-plots for HOMA-IR data comparing with standard normal distribution. It is found that HOMA-IR with Log-transformation acts more like samples of normal distribution in right figure.

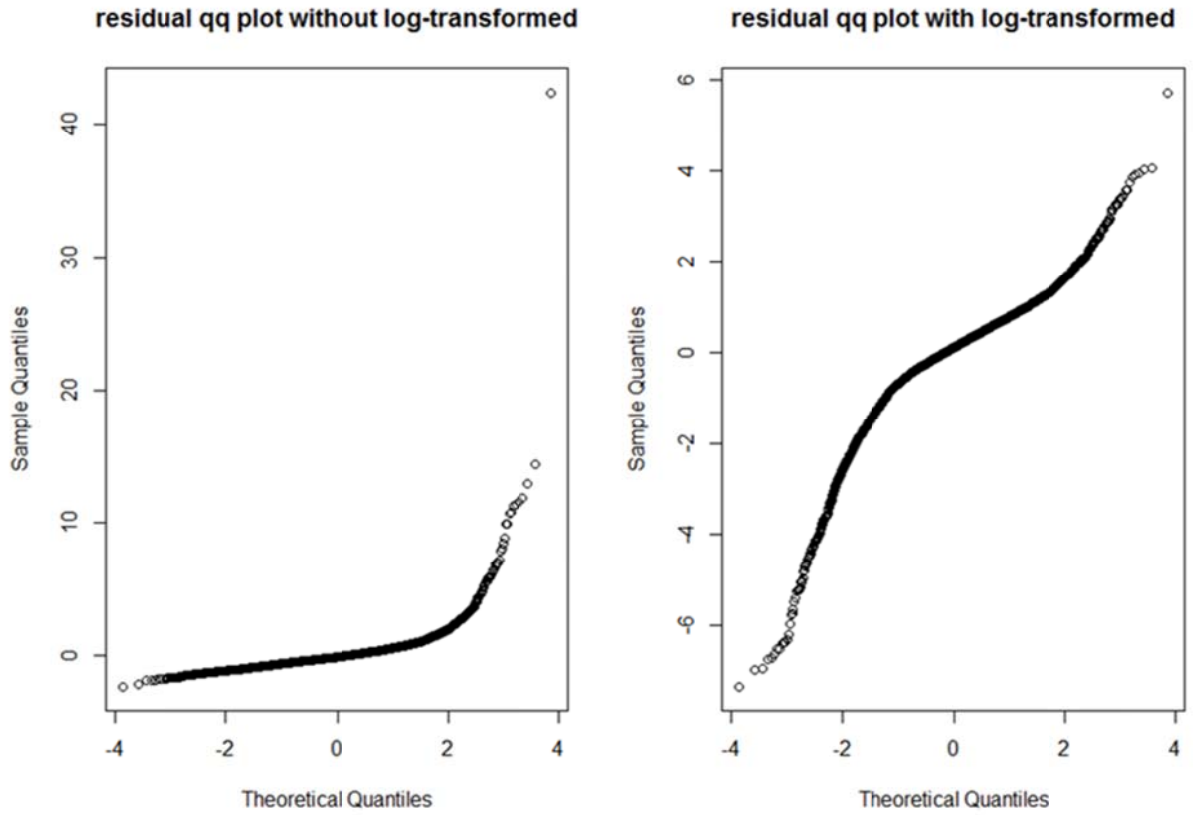


Fig 6. QQ-plot for residual analysis

We draw QQ-plot for standardized residuals. These tendencies of QQ-plots are very similar to Fig 5.

Table5 Results of real data anaylsis

	Best_Combi	CVC	Aver_Test_BA
Score based GMDR without log transformed	rs6658650 rs6494835	9	0.53256
L-estimator GMDR without log transformed	rs6658650 rs7680002	8	0.525159
M-estimator GMDR without log transformed	rs6658650 rs7680002	4	0.528727
Score based GMDR with log transformed	rs576563 rs2920792	4	0.512535
L-estimator GMDR with log transformed	rs4915657 rs7500315	9	0.518679
M-estimator GMDR with log transformed	SNP_A-2265239_A rs9353581	9	0.522633

Table 5 shows the results of several GMDR methods. If we used original HOMA-IR as response, rs6658650, rs6494835 and rs7680002 were selected by GMDR methods. The SNP rs6658650 was selected by all GMDR methods remarkably. However, SNPs which we were found by using HOMA-IR as response are no reported in researches about any relationships with any phenotypes. If we used log-transformed HOMA-IR as response, each 2 different SNPs were selected by each GMDR methods. These 6 SNPs have not been reported in the literature about any relationships with any phenotypes. However, CVC scores of L-estimator GMDR and M-estimator GMDR are higher than score-based GMDR clearly. This result coincides with our simulation results. However, it is not clear that these results suggest that L-estimator GMDR or M-estimator GMDR could find true gene-gene interactions, because two results of each GMDR methods denoted different SNP sets.

Discussion

For searching hidden heritability, it is important to find effects of gene-gene interactions efficiently. GMDR is the method which is acknowledged method for detecting gene-gene interactions. However we demonstrate with a simple example that GMDR method lose the power when erroneous samples exist. For overcoming this loss of power, we proposed L-estimator GMDR and M-estimator GMDR.

In simulation studies, we found that the score-based GMDR lose the power critically by outlier. In contrast, L-estimator GMDR and M-estimator GMDR work well when outliers existed. Table3 compares power of types of GMDR in several simulation models. L-estimator GMDR has the best performance when mixed proportion is high. It means that L-estimator has the good power when the proportion of outliers is high. In contrast, M-estimator GMDR has the best performance when mixed proportion is low or zero. This results show that L-estimator GMDR is more robust than M-estimator GMDR, however L-estimator has the worse power in the pure data than other GMDR methods. Thus M-estimator GMDR can be thought balanced method between score-based GMDR and M-estimator GMDR. However, there is no consideration about linkage disequilibrium(LD) relationships between SNPs in our simulation settings. In this reason, performances of GMDR

methods might be overestimated comparing with performances in real data.

In real data analysis, we analyzed about gene-gene interaction effects of HOMA-IR in KARE data. We could not find some important relationship between HOMA-IR and SNPs which was found by robust GMDR. However, if we performed log-transformed to HOMA-IR, versions of GMDR what we proposed were more consistent in their selections than original GMDR. In this analysis, we found that if we set threshold for M-estimator GMDR higher than some points, result from M-estimator GMDR is same as result from score-based MDR.

In further works, we will perform simulation with various LD relationships settings between SNPs for reflecting real data more precisely. We will propose the methods for finding optimal threshold of M-estimator GMDR and L-estimator GMDR for high efficiency.

References

[1] Sohee Oh, Jaehoon Lee, Min-Seok Kwon, Bruce Weir, Kyooseob Ha and Taesung Park. 2012. A novel method to identify high order gene-gene interactions in genome-wide association studies: Gene-based MDR,BMC Bioinformatics 2012, 13(Suppl 9):S5

[2] Gilbert-Diamond, D. and Moore, J. H. 2011. Analysis of Gene-Gene Interactions. Current Protocols in Human Genetics. 70:1.14.1-1.14.12.

[3] Freitas, A.A. 2001. Understanding the crucial role of attribute interaction in data mining. Artif. Intel.Rev. 16:177-199.

[4] Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH.,2001, Multifactor-dimensionality reduction reveals high-order , Am J Hum Genet. 2001 Jul;69(1):138-47. Epub 2001 Jun 11 interactions among estrogen-metabolism genes in sporadic breast cancer.

[5] Xiang-Yang Lou, Guo-Bo Chen, Lei Yan, Jennie Z. Ma, Jun Zhu, Robert C. Elston, and Ming D. Li, 2007. A Generalized Combinatorial Approach for DetectingGene-by-Gene and Gene-by-Environment Interactionswith Application to Nicotine Dependence,The American Journal of Human Genetics, Volume 80, Issue 6, 1125-1137, 1 June 2007

[6] Chung Y, Lee SY, Elston RC, Park T. 2007. Odds ratio based multifactor-dimensionality reduction method for detecting gene-gene interactions.Bioinformatics. 2007 Jan 1;23(1):71-6. Epub 2006 Nov 8.

[7] Cattaert T, Calle ML, Dudek SM, Mahachie John JM, Van Lishout F, Urrea V, Ritchie MD, Van Steen K.2011. Model-based multifactor dimensionality reduction for detecting epistasis in case-control data in the presence of noise.Ann Hum Genet. 2011 Jan;75(1):78-89.

[8] Gui J, Moore JH, Kelsey KT, Marsit CJ, Karagas MR, Andrew AS. 2011.A novel survival multifactor dimensionality reduction method for detecting gene-gene interactions with application to bladder cancer prognosis.Hum Genet. 2011 Jan;129(1):101-10

[9] Lee S, Kwon MS, Oh JM, Park T. 2012. Gene-gene interaction analysis for the survival phenotype based on the Cox model.Bioinformatics. 2012 Sep 15;28(18):i582-i588.

[10] A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. Cho YS, Go MJ, Kim YJ, Heo JY, Oh JH, Ban HJ, Yoon D, Lee MH, Kim DJ, Park M, Cha SH, Kim JW, Han BG, Min H, Ahn Y, Park MS, Han HR, Jang HY, Cho EY, Lee JE, Cho NH, Shin C, Park T, Park JW, Lee JK, Cardon L, Clarke G, McCarthy MI, Lee JY, Lee JK, Oh B, Kim HL. Nat Genet. 2009 May;41(5):527-34

[11] Gayoso-Diz P, Otero-Gonzalez A, Rodriguez-Alvarez MX, Gude F, Cadarso-Suarez C, García F, De Francisco A.2011. Insulin resistance index (HOMA-IR) levels in a general adult population: curves percentile by gender and age. The EPIRCE study.Diabetes Res Clin Pract. 2011 Oct;94(1):146-55

[12] Joint Identification of Multiple Genetic Variants via Elastic-Net Variable Selection in a Genome-Wide Association Analysis. Cho S, Kim K, Kim YJ, Lee JK, Cho YS, Lee JY, Han

BG, Kim H, Ott J, Park T.Ann Hum Genet. 2010 Sep 1;74(5):416-28

[13] Chawla, N. (2009). Mining when classes are imbalanced, rare events matter more, and errors have costs attached. SDM.

[14] Velez, D. R., White, B. C., Motsinger, A. A., Bush, W. S., Ritchie, M. D., Williams, S. M. and Moore, J. H. (2007), A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. Genet. Epidemiol., 31: 306-315.

[15] Roger Koenker & Quanshui Zhao (1994): L-estimatton for linear heteroscedastic models, Journal of Nonparametric Statistics, 3:3-4, 223-235

[16] M. Owen. Tukey's Biweight Correlation and the Breakdown. 2010. Master's thesis. Pomona College.

[17] Teri A. Manolio, Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, Erin M. Ramos, Lon R. Cardon, Aravinda Chakravarti, Judy H. Cho, Alan E. Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N. Rotimi, Montgomery Slatkin, David Valle, Alice S. Whittemore, Michael Boehnke, AndrewG. Clark, Evan E. Eichler, Greg Gibson, Jonathan L. Haines, Trudy F. C. Mackay, Steven A. McCarroll, and Peter M. Visscher. 2009. Nature. 2009 Oct 8;461(7265):747-53.

국문초록

전장 유전체 연구는 이미 수백 개의 질병 또는 형질들과 유전체들간의 연관성을 밝혀내었다. 대부분의 전장 유전체 연구에서는 질병 또는 형질들과의 연관성이 매우 큰 유전체들에만 집중하여 결과를 냈다. 이 결과 많은 형질들이 밝혀졌지만 그 형질들을 이용하여 질병의 유전과 발생을 설명에 대해 설명하기에는 부족한 점이 나타나게 되었다. 이 문제를 해결하기 위하여 많은 연구자들은 유전자와 유전자 또는 유전자와 환경간의 교호작용에 대한 연구를 시작하게 되었고 교호작용의 존재를 밝혀나가게 되었다. 2001년에 Ritchie와 그녀의 동료들은 다변량 차원 축소 기법(MDR)을 유전자와 유전자 또는 유전자와 환경간의 교호작용을 찾아내기 위해 제작하게 되었다. 다변량 차원 축소 기법은 유전자와 유전자 간의 교호작용을 해석하기 좋고 언제나 값을 만들어 낼 수 있다는 장점이 있다. 하지만 이 방법은 사례조절연구(case-control study)에서만 사용이 가능하고 환경 변수를 고려할 수 없다는 단점이 존재한다. 이런 단점을 해결하기 위하여 Xinag-Yang과 그의 동료들은 일반화 다변량 차원 축소 기법(GMDR)을 만들게 되었다. GMDR은 앞에서 언급한 MDR의 단점들을 해결할 수 있었지만 자료에 따른 결과의 변화가 매우 민감하게 나타난다는 문제점이 있다. 또는 GMDR의 경우 어떠한 예측 불가능한 관측 값을 갖는 표본의 영향을 많이 받는다. 그렇기 때문에 본 연구에서 우리는 먼저 예측 불가능한 관측 값들이 자료에 섞여 있을 때 발생하는 GMDR의 예측능력 하락을 보이고 두 가지 GMDR의 이런 문제를 해결할 것이라 기대하는 방법들을 제시하게 될 것이다. 우리가 제시하는 두 가지 방법은 L-관측량 GMDR과 M-관측량 GMDR이다. L-관측량 GMDR과 M-관측량 GMDR은 강건한 통계적 방법들을 기초로 만들어진 방법들이다. 그러므로 이 방법들이 GMDR이 가지는 자료에 민감한 문제를 해결하기에 충

분할 것이라 보고 모의실험을 제작하여 우리의 방법과 GMDR 방법을 비교해 보고 실제 자료에 적용하여 보았다. 그 결과 GMDR 방법에 비해 L-관측량 GMDR이나 M-관측량 GMDR 방법이 더 예측 불가능한 관측 값들의 영향을 줄일 수 있다는 것을 확인하였다.

Keywords : 유전자, 교호작용, MDR, GMDR, 강건성, L-관측량, M-관측량
Student Number : 2011-20241