Heritability Estimation of Dichotomous Phenotypes Using Liability Threshold Model with Ascertained family-based samples

**Abstract**

Numerous methods for estimating the heritability have been proposed. However unlike quantitative phenotypes, heritability estimation for dichotomous phenotypes suffers from computational and statistical complexity, and heritability has been limitedly utilized. In this article, we developed statistical method to estimate heritability of dichotomous phenotypes with the liability threshold model for ascertained family-based samples. Liability threshold model assumes dichotomous phenotypes are determined with unobserved latent variable which are normally distributed, and it can be applied to the general pedigree data. The proposed methods were applied to simulated data and Korean type-2 diabetes family-based samples, and accuracy of estimates by the proposed methods was compared with other methods.

**Introduction**

Phenotypes are affected by both environmental factors and genes, and family members are expected to be similar, because of their genetic similarity. Heritability was defined to quantify such phenotypic similarity attributable to heritable components, and has been widely used to understand the genetic architecture of phenotypes. For instance, heritability can be used to compare the importance of genetic components among different phenotypes. Furthermore if large-scale genetic data are available, genetic correlation matrix can be estimated, and its incorporation to linear mixed model enabled the SNP heritability estimation. SNP heritability indicates the relative proportion of variance attributable to the genotyped SNPs, and they were popularly used to identify the amount of the missing heritability.

Estimation of broad-sense heritability requires bilinear relatives such as sibling or monozygotic twins, and in practice, narrow-sense heritability has been often utilized. Narrow-sense heritability is defined as the proportion of the total phenotypic variation explained by additive genetic effects [1]. Various methods have been developed for estimating the heritability of the continuous traits. For instance, restricted maximum likelihood (REML) method based on the linear mixed model (LMM) [2-4] or polygenic score method [5] can be used for estimating the heritability of the continuous traits. For the dichotomous traits, generalized linear mixed models (GLMM) or liability threshold model (LTM) have been often utilized [6, 7]. The liability threshold model assumes there are unobserved continuous liability scores, and subjects are affected if they exceed certain threshold [8-11].

In this article, we focus on heritability estimation for dichotomous phenotypes. There are multiple factors which can bias variance estimation for dichotomous traits. In particular, family-based samples are usually ascertained with probands. Probands means family member through whom his/her family's members comes to study. Multiple literature demonstrated that such ascertainment can produce substantial bias of variance estimates [3]. For instance, if phenotypes are rare and families are randomly selected, the number of affected individuals is often very small. Therefore families are ascertained by using affected probands. In such scenario, most of their relative may be unaffected, and unless family size is large, negative correlation can be observed because probands are affected and their relatives are unaffected. Several approaches have been proposed to adjust such ascertainment bias. GCTA adjusts estimated heritabilities by assuming that the level of ascertainment bias is same among individuals [3]. However, families are ascertained with probands and the effect of ascertainment bias is heterogeneous according to their familiar relationship [12]. For instance, ascertainment bias for grandparents of the proband is expected to be around a half compared to parents.

In this article, we developed a new method to estimate Heritability based on the Liability Threshold Model for binary traits (LTMH) which can be applied to the extended pedigree structure. Using Expectation-Maximization (EM) algorithm, the proposed method jointly estimates maximum likelihood estimators (MLE) for heritability and coefficients of covariates [13]. Furthermore the proposed method maximizes the conditional likelihood conditioned on disease statuses of probands via conditional EM (CEM) algorithm [14], and ascertainment bias can be adjusted. We also developed the conditional expected score test (CEST) to test whether heritability is equal to zero. Extensive simulation studies showed that heritability estimates from the proposed methods are generally unbiased even for the ascertained family-based samples. Estimates from GCTA are unbiased for randomly selected families, but the bias becomes substantial for ascertained families. Also we found that CEST for the heritability was statistically conservative, but achieved reasonable statistical power estimates. Last we applied the proposed method to estimate the heritability of type-2 diabetes (T2D) with ascertained family-based samples of Korea and those estimates were shown to illustrate the practical value of the proposed methods.

**Methods and Materials**

**Notations and Disease Model**

We assume that there are *n* independent families and family *i* has family members (). We consider the liability threshold model, and assume dichotomous phenotypes are determined by the unobserved continuous liability score. The liability score of subject *j* in family *i* is denoted by and they are determined by summing the environmental/genetic effects, polygenic effect and random error. The covariates including environmental/genetic effects for subject *j* in family *i* is denoted by andwe assumed that covariates are standardized. In this article, we assumed there are *p* covariates. The random effects including polygenic effect and random error for subject *j* in family *i* are denoted by . The vector forms of those components for family *i* are denoted by

and .

Liability scores of family members are usually correlated, and we assumed that those are normally distributed as follows:

where . We let be the kinship coefficient matrix multiplied by two, and be the dimensional identity matrix. Under the polygenic model with additivity of genetic effects across loci and linkage equilibrium among loci, we can get

where , and are the variances of additive, dominant and environmental effects in the population, and and are the dominant genetic variance and the covariance of additive and dominant effects in the homozygous population respectively [15-17]. , and are the functions of the condensed coefficients of identity [17]. For simplicity, we assume that all variance components other than and are zero, and the sum of and is equal to one. If we denote the heritability as , then the variance-covariance matrix of is expressed by

The dichotomous phenotypes for subject *j* in family *i* is denoted by and they are coded as 1 for cases and 0 for controls. Its phenotype vector for family *i* is denoted by

.

In a liability threshold model, is determined by ; if is larger than a certain threshold value *c*, becomes 1, and otherwise it becomes 0. *c* can be determined from the prevalence of the diseases, because *c* should be the inverse of the cumulative distribution function of the prevalence. For each observed , we can infer the range of its , . For instance, if , then is bounded in , and otherwise, is bounded in . The lower and upper bounds of the liability for the family *i* are denoted by

and **.**

Based on above notations, all subjects can be expressed in the following vector forms:

and.

Under those notations, we assumed that **L** follows multivariate normal distribution with mean and variance-covariance matrix which are block diagonal matrix consisting of .

## Heritability Estimation using the EM Algorithm

The EM (Expectation-Maximization) algorithm [13] was used to estimate based on the complete data consisting of observed phenotypes, **Y**, and unobserved liabilities, **L**. The joint probability density function (pdf) of the complete data can be decomposed into the marginal pdf of **L** and the conditional pdf of **Y** given **L** which has the support of (**a, b**). It can be formulated as

.

If we denote the parameters of interest as , and the log-likelihood of the complete data will be the sum of log-likelihoods for each family, as follows,

In the E-step of the EM algorithm, the conditional expectation of **L** given **Y** was taken to the , where the estimates for the parameters of the previous iteration were used. If we assume that the *k*th iteration has been performed and denote the estimates for the parameters at the *k*th iteration as , then the conditional expectation will be

and

where and . and are equal to the first moment and the second moment of the multivariate truncated normal respectively, and R package *tmvtnorm* was utilized for calculation [18].

In M-step of EM algorithm, we maximize with respect to. Since is the concave function, we can find the maximizer by solving . The partial derivative with respect to is

and, which satisfies becomes

To emphasize that the root is the function of , it was denoted by . Unfortunately, there is no closed form of the root which , and generalized EM algorithms were applied. was updated with Newton-Raphson algorithm [19]. After we obtained the maximizer of in maximization step, we update to and repeat EM steps until convergence. The detailed algorithm is provided in the Appendix (A).

Note that is the unbiased estimator of and it can be easily proven by

assuming we got after *m* iterations [20].

**Lagrangian Multiplier and Karush-Kuhn-Tucker Condition**

Unlike , the parameter space of is restricted to , and the objective function should be maximized under the restriction as follows:

subject to .

This objective function can be maximized with the method of Lagrange multiplier [21] under the Karush-Kuhn-Trucker (KKT) conditions [22]. The constraint is equivalent to and , and by the Lagrangian multiplier, the object function becomes

where . We can find the solution that maximizes subject to by finding and satisfying the following three conditions known as KKT conditions:

1) *Stationarity* : ,

2) *Complementary slackness* : and ,

3) *Dual feasibility* : for .

More specifically, for the *Stationarity* condition, is identical to , providing that . Replacing with , we get

and it is equivalent to

Note that the left of equation is a function of , denoted by . Applying *Complemntary slackness* condition to above equation, becomes , or . If we assume and , then and it will be non-positive if the assumptions are met by the *Dual feasibility* condition. Similarly, when and are assumed, and it will be non-negative if the assumptions are satisfied. If none of them are met, and are automatically zero, and thus optimization can be done without any restrictions on . The concept is illustrated in Figure 1.

**Ascertainment Bias-corrected Heritability Estimation**

Ascertainment of each family is conducted with probands, and statistical inferences about heritability can be misled unless ascertainment is correctly adjusted. We assume the first family member in each family is a proband and the other family members are non-probands. To distinguish probands and non-probands, we added superscripts *P* and *NP* respectively, and vectors for liabilities, covariates, phenotypes and bounds of liabilities for non-probands in family *i* are denoted by

and **.**

By the same context, those for a proband in family *i* are defined as , ,, and , respectively. Liability vectors for probands and non-probands across entire families are denoted by

and ,

and vectors for other variables are also defined in the same way.

To adjust the effect of ascertainment on heritability estimates, we estimate parameters with the following conditional likelihood:

If we let , the log of the conditional likelihood is . The objective function of the EM algorithm is a global lower bound for the log-likelihood [23], and if we let the lower bound for and the upper bound for , the global lower bound can be obtained by

.

At , can be obtained by

where is the entropy. The upper bound for can be defined as [14]. Therefore, the global lower bound of the log-likelihood at becomes

We assume probands are independent with each other, and proband was randomly selected from the population with the probability . Then, is simply given by

where .

Here is formulated as a function of the cumulative distribution function of the standard normal, , by

The MLE for are obtained by iteratively maximizing the objective function until convergence, and the detailed algorithm for maximization is provided in Appendix (B).

**Conditional Expected Score Tests**

and are required to parameterize the relationship between covariates and **Y** at the unobserved liability scale, and we consider the conditional expected score test (CEST) [13, 24, 25] because

For simplicity, we assumed that the prevalence is correctly specified and samples are randomly selected. The conditional expected score based on the complete data for family *i* is

where , and . Note that and are also the function of . If we let and denote and respectively, then the score statistics can be obtained by

where , and .

The variance-covariance matrix of is calculated with the observed Fisher information matrix [26, 27]. The observed Fisher information matrix is given by

and it is equivalent to

.

Therefore if we let be the dimension of , and and are MLEs, we can provide the following statistics [26, 27]:

under .

For the test of , the likelihood is maximized at with 50% probability and at the positive real number with 50% probability under . Thus we consider

under .

**Simulation studies**

Simulation studies were conducted under two different scenarios; families were randomly selected (scenario 1) or ascertained with probands (scenario 2).

For scenario 1, 500 families were randomly generated. For scenario 2, 50,000 families for each replicate were firstly generated. Then 500 probands were selected from affected individuals, and their family members were ascertained. For both scenarios, we considered nuclear families and the number of siblings are 1, 2, 3 and 4 with proportions of 0.2, 0.3, 0.3 and 0.2 respectively. Liabilities were determined by summing a main genetic effects, polygenic effects and random errors. Sums of polygenic effects and random errors were generated from the multivariate normal distribution with heritability 0.05, 0.2 and 0.4. The main genetic effects were obtained by the product of and the number of disease alleles. Disease allele frequency was assumed to be 0.2 and genotype frequencies were obtained under the Hardy Weinberg Equilibrium. Founders’ genotypes in each family were generated from B(2, 0.2) and non-founders’ genotypes were obtained by the Mendelian transmission. was obtained by , and disease allele frequency (), with the following equation:

was assumed to be 0.005 and then was 0.1253. Once liabilities were generated, they were transformed into being affected if they were larger than the threshold *c*, and otherwise considered as being unaffected. *c* was chosen to preserve the assumed prevalences (*q*).

The performance of the proposed method was evaluated with 2,000 replicates for various combinations of and prevalences. For evaluation of statistical testing for , the prevalences (*q*) were set to be 0.1 or 0.2, and were to be 0.2 or 0.4. For evaluation of statistical testing for , we assumed *q* = 0.05, 0.1 or 0.2, and = 0, 0.2 and 0.4. All results were compared to the GCTA results for each scenario.

**Application to the Family-based Samples of Type 2 Diabetes**

The proposed method was applied to the cross-sectional study of T2D patients conducted by Seoul National University Hospital in Korea. T2D were diagnosed according to World Health Organization criteria of T2D [28]. They preferentially included T2D patients having positive family history of T2D in first-degree relatives and 681 probands were recruited. Their family histories of T2D were obtained based on memory of probands, but those who were positive for the 75-g oral glucose tolerance test among relatives were excluded from the study. We also excluded subjects who have no age information, and 4,149 non-probands including 1,115 T2D patients and 648 affected probands finally remained. For our analyses, effect of age was adjusted by including them as a covariate, and standardized age was incorporated. The prevalence of T2D was set to be 10.9% [29], and the heritability of T2D was estimated with the proposed method with adjustment of the ascertainment bias.

**Results**

**Evaluations with simulated samples**

We evaluated the accuracy of parameter estimates with simulated data. For the scenario 1, we assumed randomly selected family-based samples, and means and standard deviations (SD) of and from 2,000 replicates are shown in Table 1. The true value for is assumed to be 0.1253, and estimates for by LTMH are always close to the true values. For , estimates for LTMH and GCTA are similar if prevalence is 0.1 or 0.2 even though standard errors of estimates by LTMH are always smaller than those by GCTA. If it is 0.05 and is 0.4, bias of estimates by GCTA becomes much larger. Figure 2 shows the distribution of and both methods accurately estimates if the prevalence is large. However estimates of the GCTA are more widely distributed than those of the LTMH, and we can conclude that LTMH generally performs better.

Table 2 provides summaries for parameter estimates for ascertained families. According to the results, the most of GCTA estimates are 0 and its estimates suffer from the ascertainment bias. However estimates of and by LTMH are always close to the true values and these results show the robustness against the ascertainment bias (Table 2). Interestingly standard errors of estimates by LTMH for ascertained families are small, compared to those under the absence of ascertainment. The proportion of affected individuals is expected to be very small for rare diseases, but ascertainment of affected probands and familiar correlations increases the number of affected individuals, which may explain the smaller standard errors of heritability estimates with ascertained families. However further investigations are necessary.

We also evaluated the performance of CEST for hypothesis testing. We considered and results about empirical sizes are shown in Table 3. It should be noted that ascertainment does not affect the statistical inferences if . Results show that LTMH were slightly conservative if *q* = 0.05 or 0.2, but its type-1 error estimates are very close to the nominal significance levels if *q* = 0.1. This conservativeness may indicate the overestimate of the variance. Table 3 also shows the statistical power estimates. We assumed that the true is 0.2 or 0.4, and *q* is 0.05, 0.1 and 0.2. The statistical power estimates increases as the true heritability, prevalence or both increases, and very large empirical power estimates were obtained for the larger prevalence. We also evaluated the statistical performance of the score tests for (Tables 4). It shows that the score tests for are not conservative and always preserve the nominal significance level under the null hypothesis, Empirical power estimates for were assessed using 2,000 replicates at several significance levels and they increases as the prevalence, heritability or both get larger.

**Applications of LTMH and CEST to Type 2 Diabetes**

To evaluate the performance of LTMH with real data, we considered the family-based samples of T2D dataset. Table 5 shows their descriptive statistics [30]. There were 1,736 T2D patients (36.75%) and average age is 48.63 years old with SD of 15.7 (Table 5). Proportions of males and females are similar. All non-probands are the first-degree relatives of probands, and the familiar relationship with the largest number is siblings (59.22%), followed by offspring (32.85%).

LTMH was applied to the family-based samples of T2D dataset, and heritability of the T2D was estimated. Estimated heritability of T2D was 29.44%, and it was statistically significant under the significance level of 0.05 (P-value = 1.2010-5). This result is slightly overestimated in comparison to heritability estimates of T2D (26%) using ACE model based on the twin data [31]. This difference may be attributable to the racial differences. Coefficient estimates for unstandardized age was 0.051 (0.8 for standardized age), which means that the threshold underlying disease is reduced by 0.051 at the liability scale if the age increases by 1. The function of age is well described in Figure 3A which illustrates the probability of being affected with the T2D according to age. Results show that the risk increases monotonically by the age reflecting the reduction effect on the threshold and individuals with more affected relatives with the T2D have greater risk. In comparison to random sample, the influence of family history is greater at young age, providing familial risk for early-onset T2D is highly important (Figure 3B).

**Discussion**

In this article, we proposed a new method to estimate a heritability on dichotomous trait based on the liability threshold model for the ascertained family-based samples. A simulation study demonstrated that LTMH generally provides more accurate estimates of heritabilities than the GCTA, and their differences are substantial for ascertained families. To our knowledge, there is no method to deal with the ascertained samples in estimating heritability on dichotomous trait. Furthermore, we assessed the statistical performance of the CEST. Statistical power estimates were evaluated for various situations and substantial power improvement was observed for the common disease rather than the rare disease.

In spite of the power improvement of the proposed methods, there are some limitations. First, the CEST for was conservative. We found that the likelihood for is not symmetric under the null hypothesis and it may be attributable to the misspecified weights for and for the distribution of the CEST under . Fortunately, we found that such inflation does not affect the statistical power but some modification such as bootstrapping is necessary. Second, the proposed method is the computationally very intensive when the family size, , is large, and the expected computational time is proportional to . The most computational burden is from the calculation of conditional expectation in the E-step of EM algorithm. The computational burden can be reduced by reducing the number of iterations for the EM algorithm or approximating the moment of the multivariate truncated normal. The formal can be achieved using EM acceleration methods which can make EM dramatically faster such as Aitken acceleration, conjugate gradient acceleration, quasi-Newtonian acceleration and parameter expansion acceleration [32-36]. For the latter, conditional expectation may be approximated with some numerical algorithm such as Laplace approximation. These will be investigated as our further research.

Heritability has been importantly utilized for genetic epidemiology. However its estimation for dichotomous phenotypes has been very complicated because of ascertainment bias. In spite of several limitations, the proposed method enabled heritability estimation of dichotomous trait with ascertained families, and may be a promising strategy to estimate narrow-sense heritability for various diseases. LTMH is implemented in R language and source codes is freely available at <http://healthstat.snu.ac.kr>.

**Table 1. Accuracy of and**  **for randomly selected families (scenario 1)**. Parameter estimates from 2,000 replicates were summarized using mean (top) and standard error (bottom). The true value of is 0.1253. SD stands for the standard deviation.

**Table 2. Accuracy of and**  **for ascertained families (scenario 2)** Parameter estimates from 2,000 replicates were summarized using mean (top) and standard error (bottom). The true value of is 0.1253.

**Table 3. Type-1 error and power estimates of the proposed test for under scenario 1**. The empirical sizes () and powers (.2 and 0.4) were estimated with 2,000 replicates at three significance levels. We considered prevalence of 0.05, 0.1 and 0.2.

**Table 4. Type-1 error and power estimates of the proposed test for**  **under scenario 1**. The empirical sizes () and powers (.005) were estimated with 2,000 replicates at three significance levels. We considered heritability of 0.2 and 0.4, and prevalence of 0.1 and 0.2.

**Table 5. Demographic characteristics for study participants.** For categorical variables, number of subjects and their proportions are provided. For continuous variables, means and standard deviations are provided.

**Figure 1. Illustration of KKT condition using a toy example.** The exemplary concave function, , was created to show how to find the optimal value which maximizes in the parameter space. The parameter, , can be between zero and one, and its parameter space is grayed out. (A) If the value which maximizes is negative, the tangent slopes at both zero and one will be negative. It violates the KKT condition that tangent slope is negative at one. However, the negative tangent slop at zero satisfies the KKT condition, so the maximizer within the parameter space is zero. (B) When the value which maximizes is greater than 1, the optimal value is one since positive tangent slope at one meets the KKT condition. (C) When the maximizer is located in the parameter space, tangent slopes at both boundaries of the parameter space do not satisfy the KKT condition. Therefore, restrictions do not affect the result of optimization.

**Figure 2. Boxplots for for randomly selected families (scenario 1).** True heritability was 0.05 (top), 0.2 (middle) and 0.4 (bottom), and were indicated as gray dashed line.

**Figure 3. Estimation of risks for T2D according to ages.** For a certain individual, we assume that he/she has two parents and one younger sibling, and the risk of T2D development was calculated as a function of his/her age and the number of their affected family members. X axis indicates the individual’s age, and ages of his/her father and mother were assumed to be 29 years older. The younger sibling was assumed to be 3 years younger than the participant. and the coefficient of unstandardized age were set to be 0.2944 and 0.051 respectively. (A) Probability of the participant being affected according to the number of affected family members, and (B) relative risks of being affected according to the number of affected family members.

**Appendix (A)**

**Numerical analysis for optimization of the heritability in M-step of EM algorithm**

The first derivative of with respect to is given by

where . Then, the objective function becomes

Similarly, we can get the first derivative of with respect to as follows,

where and

Finally, is updated according to the following iterative steps with and which was calculated at the previous E-step,

**Appendix (B)**

**Numerical analysis for maximizing the global lower bound**

If we denote the global lower bound for the conditional log-likelihood as , then the first derivative of with respect to ,,is given by

Here it should be noted that is equivalent to the equations (1) and (2) in the Method and Appendix (A). Using the chain rule, we can easily obtain the first derivative of with respect to as follows,

where is the probability density function for the standard normal. To apply Newton-Raphson algorithm for the objective function , we derive the first derivative of with respect to , , as follows,

and each term is given by

and

With these terms, we iteratively update using following equation until convergence,