Heritability Estimation of Dichotomous Phenotypes Using Liability Threshold Model

Wonji Kim1, Sungho Won1,2\*

1 Interdisciplinary program in Bioinformatics, Seoul National University, Korea

2Department of Public Health Science, Seoul National University, Korea

\*To whom correspondence should be addressed.

**Abstract**

Numerous methods for estimating the heritability have been proposed. However unlike quantitative phenotypes, parameter estimation for dichotomous phenotypes suffers from computational and statistical complexity and heritability has been limitedly estimated. In this article, we developed statistical method to estimate heritability of dichotomous phenotypes with the liability threshold model. 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 accuracy of estimates by the proposed methods was compared with others.

**Keywords**

Heritability, Binary trait, Score Test

**Introduction**

Heritability is an important measure for understanding the genetic architecture of the complex traits. Narrow-sense heritability is usually defined as the proportion of the total phenotypic variation explained by additive genetic effects [1]. Various methods for estimating the heritability of the continuous traits have been developed, for example restricted maximum likelihood (REML) method based on the linear mixed model (LMM) [2-4] or polygenic score method [5]. They enabled us to make a number of achievements including the conceptual approach of the “missing heritability” by accurately estimating the total heritability. For the dichotomous traits, generalized linear mixed models (GLMM) are often used to estimate genetic variance components [6, 7]. However, in GLMM based approaches, calculation of heritability is hindered since the phenotypic variance and the genetic variance are not on the same scale. Meanwhile, a liability threshold model (LTM) is usually assumed for dichotomous traits, i.e., subjects are affected if unobserved continuous liabilities following a multivariate normal distribution exceed certain threshold underlying a disease. LTM is more straightforward to estimate and interpret heritability since both the variance for environmental factors and the variance for genetic factors are defined on a liability scale [8-11]. However, most of methods do not consider familial aggregation caused by recruitment of so-called probands, which hinders their practical use.

Familial aggregation occurs when family members are entered to the study due to the subjects which have been recruited for the case-control studies or only-case studies. These family members have different characteristics from randomly selected subjects in that the probability of being included to the study is not equal to that for the randomly selected subjects, that is, the joint likelihood of liabilities for the family members are conditioned on the disease status of the proband. It has to note that a proband ascertainment is a different context with a case-control ascertainment in that all cases and controls are ascertained. GCTA which is the one of well-established algorithm for estimating heritability can adjust case-control ascertainment by using liability threshold properties but adjustment for proband ascertainment is not offered [3].

In this study, we developed a heritability estimation algorithm that can be applied to general pedigree structure based on the liability threshold model for binary traits. Using Expectation-Maximization (EM) algorithm, maximum likelihood estimators for heritability and coefficients of covariates are obtained [12]. In case of proband ascertainment, we estimate parameters to maximize the conditional likelihood given by disease statuses of probands via conditional EM (CEM) algorithm [13]. Simulation results showed that the proposed method provides stable estimation except for the scenario of low prevalence and low heritability. In comparison to the GCTA, the estimates of the proposed method are slightly biased than that of GCTA, but showed smaller variation. We applied the proposed method to the healthy TWIN study of Korea [14] and estimated the heritability of the obesity.

We also developed the Conditional Expected Score Test (CEST) which is a Rao’s score test which can infer the null hypothesis that heritability is zero. It can also apply to the hypothesis for coefficients of the covariates. In the simulation studies, CEST for the heritability seems somehow conservative in terms of small empirical sizes for the heritability. Nevertheless, statistical power seems still high when the null hypothesis is not true. For the coefficients of the covariates, type-I error seem to be well preserved and statistically very powerful. We conducted genome-wide association studies for Lymphangioleiomyomatosis (LAM) data using CEST, ordinary logistic regression (LR) and conditional logistic regression (CLR). We identified genome-wide significant SNPs and estimate heritability and effect sizes.

**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 dichotomous phenotypes are determined by the unobserved continuous liability score. The liability scores 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 which is standardized for each covariate. 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 form of each component for family *i* are denoted by

and **.**

Liability scores of family members are often correlated, and we consider the simple additive polygenic effect model. We assumed that liability scores are normally distributed as follows:

where . 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, dominance and environmental effects in the population, and and are the dominance genetic variance and the covariance of additive and dominance effects in the homozygous population respectively [18-20]. is the kinship coefficient matrix multiplied by two, , and are the functions of the condensed coefficients of identity [20] and is the dimensional identity matrix. 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 which is defined as , then the variance-covariance matrix of is expressed as follows,

.

The dichotomous phenotypes for subject *j* in family *i* is denoted by and they are coded as 1 for cases and 0 for controls. In a liability threshold model, is determined by ; if is above a certain threshold value *c*, becomes 1, and otherwise it becomes 0. *c* can be determined from the prevalence of the diseases, and the phenotype vector for family *i* is denoted by

.

Several algorithms have been suggested to estimate *c* with prevalence, *q*, known a *priori*, and in this study we used cumulative distribution function of standard normal. Therefore, once is observed, we can infer the range of the liability, , which are the lower and upper bound of respectively. For instance, if is equal to zero, 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 individuals can be expressed by following vector forms:

and.

The liability vector **L** follows multivariate normal distribution with mean and variance-covariance matrix which are block diagonal matrix consisting of .

## Estimation of the Heritability using the EM Algorithm

The EM (Expectation-Maximization) algorithm [21] 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 assumed that the *k*th iteration has been performed, the conditional expectation will be

and

where and .

In M-step of EM algorithm, we maximize with respect to. Since is the concave function, we can find the maximizer by solving the equation that the first derivative of is equal to **0**. The partial derivative with respect to is

and, which satisfies will be

To emphasize that the root is the function of , it was denoted by . Unfortunately, there is no closed form of the root which . In this process, we need a numerical algorithm to find root and Newton-Raphson algorithm [22] was utilized here. The detailed algorithm is provided in the Appendix (A).

After we obtained the maximizer of in maximization step, we updated to and repeat EM steps until convergence. When the parameter converges, we finally obtained  and . Note that is the unbiased estimator of and it can be easily proven by using that assuming we got after *m* iterations.

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

Unlike , the parameter space of is restricted to . Thus, should be estimated in with the following objective function at the *k*th iteration

subject to .

We can optimize the above objective function via the method of Lagrange multiplier [23] under the Karush-Kuhn-Trucker (KKT) conditions [24]. Changing the constraint to and , the Lagrangian, , is denoted by

where . We can find the solution that maximizes subject to by finding and satisfying the following three conditions known as KKT conditions. The first KKT condition known as *Stationarity* is

Since does not depend on the constraint, is identical to , providing that . Replacing to , we get

and it is equivalence to

Note that the left hand side is the function of , denoted by . The second condition which is called as *Complementary slackness* is and , and is equivalent to

.

Since and , cannot be satisfied simultaneously, will be , or . Of them, and are related to the last condition, *Dual feasibility*, which is for . 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 the problem becomes the optimization with no restrictions on since the constraints have no meaning. The concept is illustrated in Figure 1.

**Robust Estimator to the Misspecified Prevalence**

We showed that is the unbiased estimator regardless the choices of the heritability and the prevalence. On the other hand, the estimator for the heritability is biased and the prevalence is involved in   in such a complicated manner. In our optimization procedure, the only part where the prevalence is involved in is the E-step of EM algorithm, which takes the conditional expectation to the liabilities. Performing optimization by setting the prevalence as a parameter is too complex in terms of calculating the conditional expectation of multivariate truncated normal distribution. Alternatively, if the true prevalence is lower than the assumed prevalence, it can be corrected by shifting liabilities to the right as a whole and it is achievable quite simply by including a positive intercept. This concept is equally applicable in the opposite situation. Therefore, inclusion of the intercept in the parameters can have the effect of correcting the misspecified prevalence. We denoted the parameter for the intercept as .

If we let and where is the dimensional vector which elements are all one, we will get the estimates for and at the *k*th iteration of the EM algorithm as follows,

where , , and =.

**Conditional Expected Score Tests**

Since and are the parameters on the liability scale, statistical significance of the estimators cannot be directly tested with the score test based on the observed data, **Y**. Alternatively, the score based on **Y** can be derived using the fact that the Fisher score based on the incomplete data is equal to the conditional expectation of the score based on the complete data given the observed data [21, 25] which can be formulated in our cases as follows,

The score test with the above statistics is also called Conditional Expected Score Test (CEST) in that it takes the conditional expectation to the score based on the complete data [26]. For simplicity, we assumed the prevalence is correctly specified.

The conditional expected score based on the complete data for family *i*, is

where , and . Note that and are also the function of . The conditional expected score statistics can be calculated by adding up for entire families, that is, .

**Simulation studies**

We generated 10,000 families with three, four, five and six members consisting of a parents and offsprings at the rates of 0.2, 0.3, 0.3 and 0.2 respectively. Liability was assumed to be affected by a main genetic factor with minor allele frequency 0.2 under the Hardy Weinberg Equilibrium, and no environmental effect. Founders’ genotypes in each family were generated from binomial distribution with the number of trials of 2 with the probability for success of 0.2, and the non-founders’ genotypes were obtained by randomly generated by the rule of Mendelian transmissions. The random effects including the polygenic effects and random errors were generated from the multivariate normal distribution with heritability 0.2 and 0.4. Finally, liability is calculated by summing the product of and the number of minor alleles and random effects. Here is obtained by the assumed the relative proportion of variance explained by the SNP (), and minor allele frequency (), with the following equation:

Here, was assumed to be 0.005 and corresponded to 0.1253. Once liabilities were generated, they were transformed into affected if larger than the threshold *c*, and otherwise were considered unaffected. The value of *c* was chosen to preserve the assumed prevalences of *q* = 0.1 or *q* = 0.2.

For each setting of and *q*, simulation was repeated 300 times with 500 families randomly selected among 10,000 families. Simulation studies were conducted according to the following scheme. First, with correctly specified prevalence, we estimated and using the basic estimation method. Second, assuming that the prevalence was 0.1 larger than true value, and were estimated with and without the tuning parameter .

**CRC Dataset**

**Results**

**Results of Simulation Studies**

For the first scheme of the simulation study which assumes prevalence correctly, some descriptive statistics of and are shown in Table 1. The estimates appear to be slightly different from true values but the bias can be reduced by increasing the family size. The estimates are concentrated around the true value even though they seem to be skewed slightly (Figure 2).

When the prevalence is misspecified, the basic estimation method provided seriously biased estimates for (Figure 3A). Interestingly, when the true prevalences were 0.1 and 0.2, was overestimated to be about 0.4 and 0.2 larger than the true respectively. According to the results, the basic estimation method seems sensitive to the specification of the prevalence and bias depends on not only the difference between true prevalence and assumed prevalence, but the true prevalence itself. Overestimation of is not surprising because observing a lower proportion of affected subjects than the assumed prevalence implies a high correlation between subjects. This phenomenon would appear to be stronger in rare disease. Whereas, was estimated to be relatively stable compared to but somewhat underestimated (Figure 3B).

In Figure 4, we displayed the distributions of estimates using the robust estimation method when the prevalence is misspecified. The bias in based on the basic estimation method seems to be drastically reduced (Figure 4A) and s are well distributed around the true value (Figure 4B). The tuning parameter, , was estimated to have negative values, which reduces the prevalence by shifting the liabilities to the right (Figure 4C).

1. Wright, S., *Evolution and the genetics of populations, volume 3: experimental results and evolutionary deductions*. Vol. 3. 1984: University of Chicago press.

2. Dempster, E.R. and I.M. Lerner, *Heritability of threshold characters.* Genetics, 1950. **35**(2): p. 212-236.

3. Gianola, D., *Heritability of polychotomous characters.* Genetics, 1979. **93**(4): p. 1051-1055.

4. Van Vleck, L., *Estimation of heritability of threshold characters.* Journal of Dairy Science, 1972. **55**(2): p. 218-225.

5. Quaas, R. and L. Van Vleck, *Categorical trait sire evaluation by best linear unbiased prediction of future progeny category frequencies.* Biometrics, 1980: p. 117-122.

6. Harville, D.A. and R.W. Mee, *A mixed-model procedure for analyzing ordered categorical data.* Biometrics, 1984: p. 393-408.

7. Foulley, J., D. Gianola, and S. Im, *Genetic evaluation of traits distributed as Poisson-binomial with reference to reproductive characters.* Theoretical and Applied Genetics, 1987. **73**(6): p. 870-877.

8. Simianer, H. and L. Schaeffer, *Estimation of covariance components between one continuous and one binary trait.* Genetics Selection Evolution, 1989. **21**(3): p. 303.

9. Gianola, D., *Genetic evaluation of animals for traits with categorical responses.* Journal of Animal Science, 1980. **51**(6): p. 1272-1276.

10. Stiratelli, R., N. Laird, and J.H. Ware, *Random-effects models for serial observations with binary response.* Biometrics, 1984: p. 961-971.

11. Landis, J.R. and G.G. Koch, *An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers.* Biometrics, 1977: p. 363-374.

12. Duffy, D.E. and T.J. Santner, *On the small sample properties of norm-restricted maximum likelihood estimators for logistic regression models.* Communications in Statistics-Theory and Methods, 1989. **18**(3): p. 959-980.

13. McCullagh, P., *Generalized linear models.* European Journal of Operational Research, 1984. **16**(3): p. 285-292.

14. Searle, S.R., G. Casella, and C.E. McCulloch, *Variance components*. Vol. 391. 2009: John Wiley & Sons.

15. Magnussen, S. and A. Kremer, *The beta-binomial model for estimating heritabilities of binary traits.* Theoretical and applied genetics, 1995. **91**(3): p. 544-552.

16. Paul, A.K. and V. Bhatia, *Modification of beta-binomial method of estimation of heritability of stayability.* J. Ind. Soc. Agril. Statist, 2001. **54**(3): p. 385-395.

17. Davies, S.W., et al., *The design and analysis of binary variable traits in common garden genetic experiments of highly fecund species to assess heritability.* bioRxiv, 2015: p. 018044.

18. Fisher, R.A., *XV.—The correlation between relatives on the supposition of Mendelian inheritance.* Earth and Environmental Science Transactions of the Royal Society of Edinburgh, 1919. **52**(2): p. 399-433.

19. Abney, M., M.S. McPeek, and C. Ober, *Estimation of variance components of quantitative traits in inbred populations.* The American Journal of Human Genetics, 2000. **66**(2): p. 629-650.

20. Jacquard, A., *The genetic structure of populations*. Vol. 5. 2012: Springer Science & Business Media.

21. Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the royal statistical society. Series B (methodological), 1977: p. 1-38.

22. Atkinson, K.E., *An introduction to numerical analysis*. 2008: John Wiley & Sons.

23. Bertsekas, D.P., *Constrained optimization and Lagrange multiplier methods*. 2014: Academic press.

24. Kuhn, H.W. and A.W. Tucker, *Nonlinear programming*, in *Traces and emergence of nonlinear programming*. 2014, Springer. p. 247-258.

25. Fisher, R.A. *Theory of statistical estimation*. in *Mathematical Proceedings of the Cambridge Philosophical Society*. 1925. Cambridge University Press.

26. Finkelstein, D.M., et al., *A score test for association of a longitudinal marker and an event with missing data.* Biometrics, 2010. **66**(3): p. 726-732.

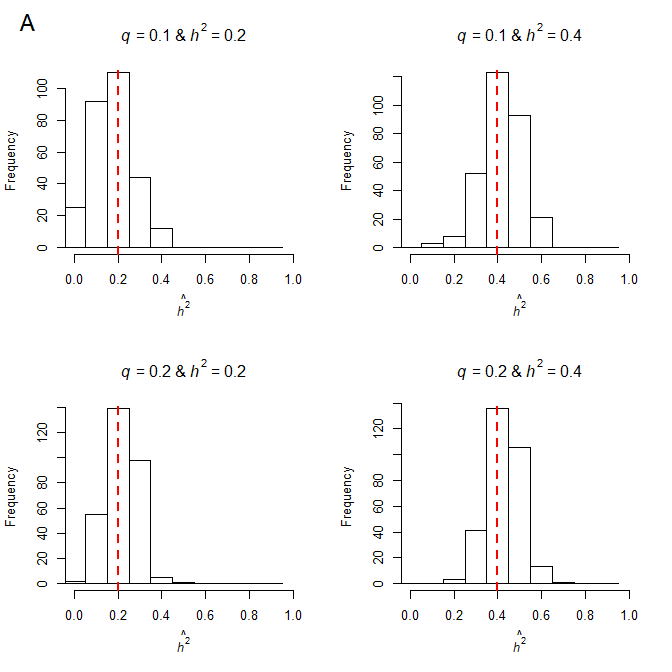
**Table 1. The Descriptive statistics for and**  **based on 300 replicates when the prevalence was correctly specified**. The estimates from 300 replicates were summarized using some descriptive statistics; Mean : observed mean of the 300 estimates, Median : observed median of the 300 estimates, SD : observed standard deviation of the 300 estimates. The true value for is 0.1253.

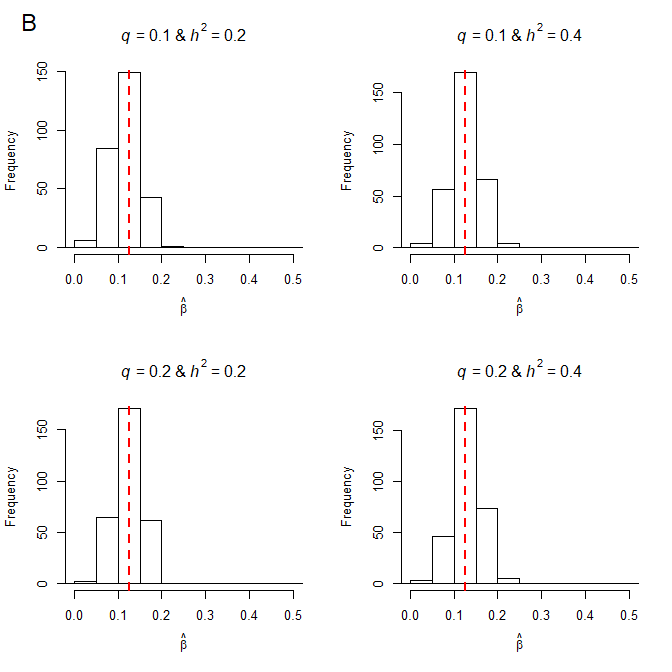
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Prevalence | Heritability | Parameter | Mean | Median | SD |
| 0.1 | 0.2 |  | 0.1146 | 0.1127 | 0.0333 |
|  | 0.1757 | 0.1684 | 0.0917 |
| 0.4 |  | 0.1244 | 0.1241 | 0.0322 |
|  | 0.4221 | 0.4232 | 0.0960 |
| 0.2 | 0.2 |  | 0.1230 | 0.1208 | 0.0300 |
|  | 0.2170 | 0.2183 | 0.0725 |
| 0.4 |  | 0.1311 | 0.1323 | 0.0317 |
|  | 0.4285 | 0.4335 | 0.0774 |

**Figure 1. Illustration of KKT condition using a toy example.** Virtual concave functions, , were created to show how to find the optimal value which maximizes in the parameter space. The parameter, , is restricted to be between zero and one, and the 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 is in violation of 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.

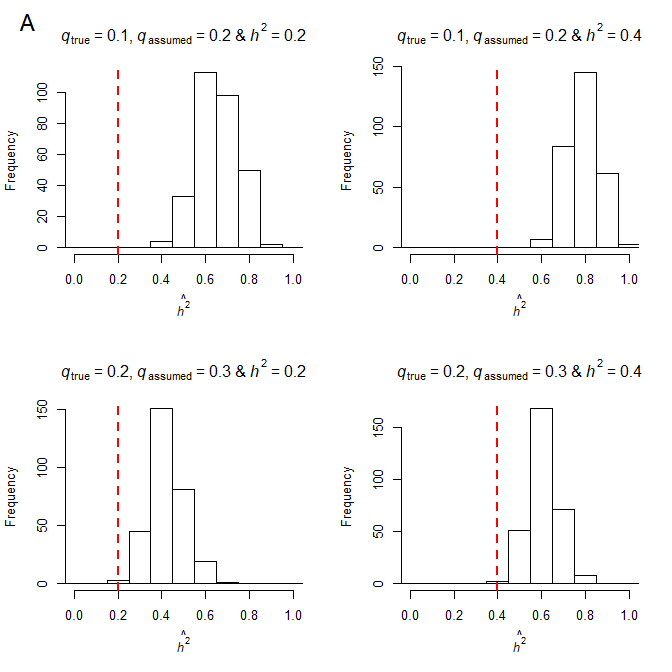
****

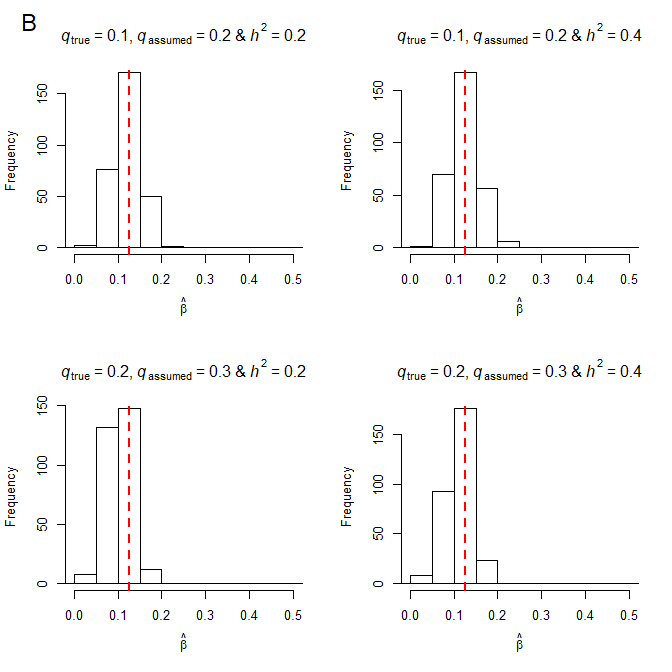
**Figure 2. Distribution of estimates based on 300 replicates when the prevalence was correctly specified.** True prevalence was 0.1 in the upper two figures and 0.2 in the lower two figures. True heritability was 0.2 in the left two figures and 0.4 in the right two figures. (A) The distribution of . Red dashed line indicates true value of . (B) The distribution of . Red dashed line indicates true value of .

****

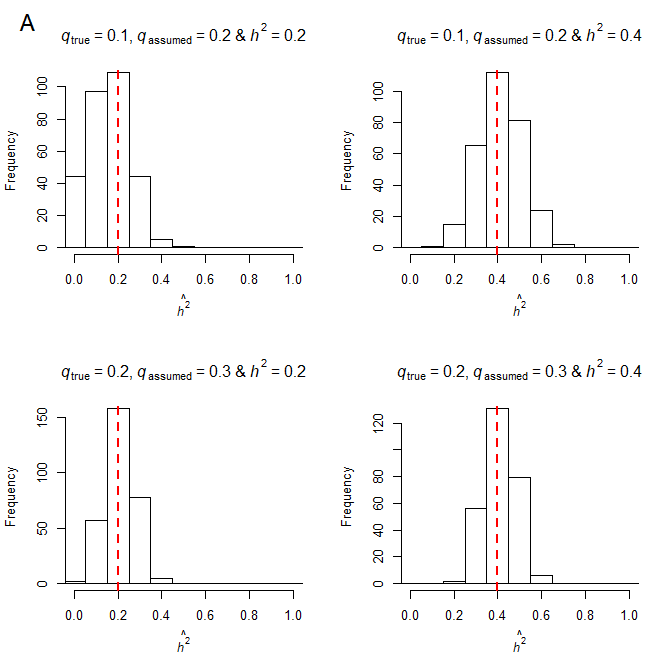
****

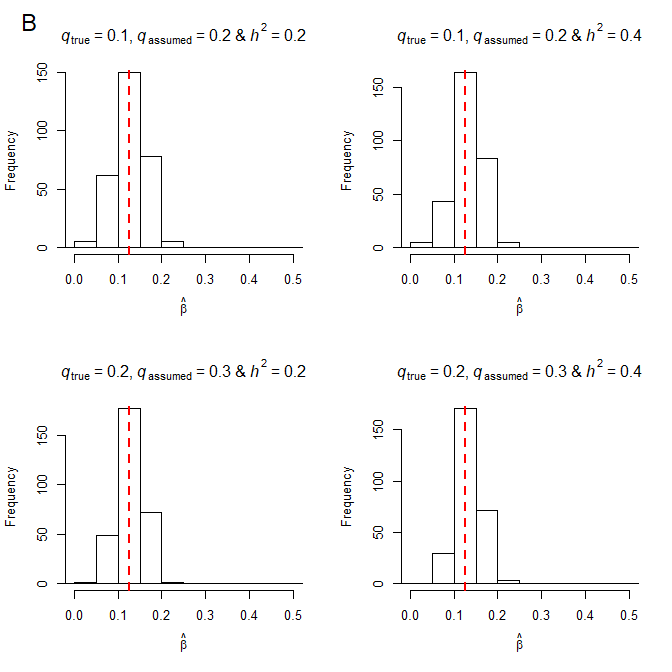
**Figure 3. Distribution of estimates without the intercept based on 300 replicates when the prevalence was misspecified.** The prevalence was assumed to be 0.1 larger than the true prevalence. True prevalence was 0.1 in the upper two figures and 0.2 in the lower two figures. True heritability was 0.2 in the left two figures and 0.4 in the right two figures. (A) The distribution of . Red dashed line indicates true value of . (B) The distribution of . Red dashed line indicates true value of .

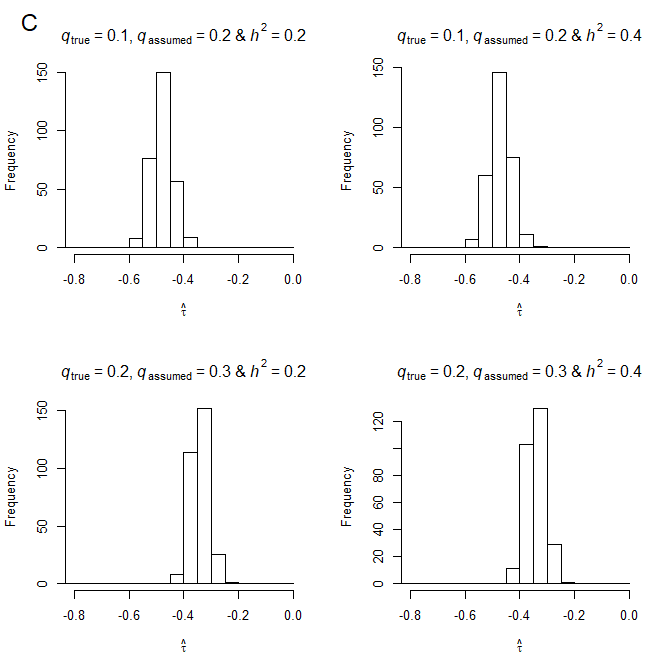
****

****

**Figure 4. Distribution of estimates with the intercept based on 300 replicates when the prevalence was misspecified.** The prevalence was assumed to be 0.1 larger than the true prevalence. True prevalence was 0.1 in the upper two figures and 0.2 in the lower two figures. True heritability was 0.2 in the left two figures and 0.4 in the right two figures. (A) The distribution of . Red dashed line indicates true value of . (B) The distribution of . Red dashed line indicates true value of . (C) The distribution of .





****

**Appendix**

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