Conference on Applied Statistics in Agriculture 2000 - 12th Annual Conference Proceedings

POINT ESTIMATORS OF HERITABILITY BASED ON CONFIDENCE INTERVALS: A CLOSED-FORM APPROXIMATION TO THE REML ESTIMATOR

Brent D. Burch

Ian R. Harris

Recommended Citation

Burch, Brent D. and Harris, Ian R. (2000). "POINT ESTIMATORS OF HERITABILITY BASED ON CONFIDENCE INTERVALS: A CLOSED-FORM APPROXIMATION TO THE REML ESTIMATOR," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1243

POINT ESTIMATORS OF HERITABILITY BASED ON CONFIDENCE INTERVALS: A CLOSED-FORM APPROXIMATION TO THE REML ESTIMATOR

Brent D. Burch and Ian R. Harris

Department of Mathematics and Statistics, Northern Arizona University,
Flagstaff, Arizona 86011, U.S.A.

ABSTRACT

Estimating heritability, the proportion of variation in phenotypic values due to (additive) genetic effects, is an important subject matter to plant and animal breeders alike. In most applications there is not an analytic expression for the restricted maximum likelihood (REML) estimator of heritability since it is obtained via an iterative procedure. The focus of this paper is to find a closed-form approximation to the REML estimator of heritability for those scenarios in which mixed linear models having two variance components are appropriate. This procedure is equivalent to constructing approximate pivotal quantities and thus confidence intervals for heritability. See Burch and Iyer (1997) and Harris and Burch (2000) for more details concerning this approach. The closed-form estimator is compared to the REML estimator by evaluating their asymptotic standard errors. An application involving yearling bulls from a Red Angus seed stock herd suggests that the closed-form estimator mimics the REML estimator and is a viable candidate for investigators seeking a non-iterative method to estimate heritability.

1 Introduction

The general mixed linear model under consideration is

$$\mathbf{Y} \;=\; \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{1}$$

where $\mathbf{Y}$ is a $n \times 1$ vector of observable phenotypic values, $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters that model environmental influences on the phenotypic values, $\mathbf{u}$ is a $m \times 1$ vector of unobservable variables representing the (additive) genetic influences on the phenotypic values, and $\mathbf{e}$ is a $n \times 1$ vector of unobservable variables representing the influence of other environmental and genetic effects on the phenotypic values that are not accounted for in the first two terms on the right side of equation (1). The matrices $\mathbf{X}$ and $\mathbf{Z}$ are known and without loss of generality, $\text{rank}(\mathbf{X}) = p$.

For example, in the application to be discussed later in the paper, $\mathbf{Y}$ is the set of measurements of loineye muscle area (measured in square inches) of yearling bulls from a Red Angus seed stock herd, $\boldsymbol{\beta}$ keeps track of the age-group of the dam of each animal, $\mathbf{u}$ refers to the (additive) genetic effect of the animals on the loineye muscle area, and $\mathbf{e}$

represents the other influences on loineye muscle area that have not been accounted for by the age of dam and the (additive) genetic effects.

In the usual manner it is assumed that $\mathbf{u}$ and $\mathbf{e}$ are multivariate normally distributed. Specifically, $\mathbf{u}$ and $\mathbf{e}$ are independent where $\mathbf{u} \sim N(\mathbf{0}, \sigma_1^2\mathbf{A})$ and $\mathbf{e} \sim N(\mathbf{0}, \sigma_2^2\mathbf{I_n})$. The variance components σ_1^2 and σ_2^2 represent the variation in the phenotypic values that are attributed to (additive) genetic effects and "other" effects, respectively. Using the distributional assumptions given above it follows that $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma_2^2\mathbf{I_n} + \sigma_1^2\mathbf{ZAZ}')$. The known matrix $\mathbf{A}$ is referred to as the relationship matrix since it describes the degree to which the animals are genetically related. Animals that are genetically related may exhibit somewhat similar physical traits.

Since σ_1^2 and σ_2^2 denote the two variance components, the quantity $\rho = \sigma_1^2/(\sigma_1^2 + \sigma_2^2)$ is the proportion of the total variation in the phenotypic values due to (additive) genetic effects. ρ is referred to as heritability and measures the degree of resemblance between relatives. Since we our focusing our attention on the variance components as well as a particular function of variance components using model (1), it makes sense to find out what functions of the data contain information about the variance components. To find REML estimators of variance components or heritability, one considers the maximization of a restricted likelihood function. The restricted likelihood function considered here is based on quadratic forms of the data which we now discuss. See Burch and Iyer (1997) for additional details.

Let $\mathbf{H}$ be a $n \times (n-p)$ matrix whose columns span the space orthogonal to the space spanned by the columns of $\mathbf{X}$ and satisfies $\mathbf{H}'\mathbf{H} = \mathbf{I_{n-p}}$. Then $\mathbf{H}'\mathbf{Y} \sim N(\mathbf{0}, \sigma_2^2\mathbf{I_{n-p}} + \sigma_1^2\mathbf{H}'\mathbf{ZAZ}'\mathbf{H})$. Note that $\mathbf{H}'\mathbf{Y}$ is a $n-p$ dimensional vector whose distribution does not depend on the $\boldsymbol{\beta}$, the vector of location parameters. Since the REML estimation procedure maximizes that part of the likelihood function which is invariant to fixed effects, one may consider using the likelihood function of $\mathbf{H}'\mathbf{Y}$ to find the REML estimator of ρ. In essence, REML estimation is a maximum likelihood procedure based on linear combinations of the data rather than the data themselves. See Harville (1977) for a detailed discussion of maximum likelihood approaches to variance component estimation.

Let $0 \leq \Delta_1 < ... < \Delta_d$ be the distinct eigenvalues of $\mathbf{H}'\mathbf{ZAZ}'\mathbf{H}$ having multiplicities $r_1, ..., r_d$, respectively. There exists an $(n-p) \times (n-p)$ orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}'(\mathbf{H}'\mathbf{ZAZ}'\mathbf{H})\mathbf{P} = Diag(\Delta_1, .., \Delta_1, ..., \Delta_d, ..., \Delta_d)$ where each Δ_i is repeated r_i times, $i = 1, ..., d$. It follows that $\mathbf{H}'\mathbf{ZAZ}'\mathbf{H} = \sum_{i=1}^{d} \Delta_i \mathbf{P}_i\mathbf{P}_i'$ where $\mathbf{P} = [\mathbf{P}_1, ..., \mathbf{P}_d]$ and each matrix $\mathbf{P}_i$ corresponding to Δ_i is of size $(n-p) \times r_i$. For $i = 1, ..., d$, $\mathbf{P_i'}\mathbf{H}'\mathbf{Y} \sim N(\mathbf{0}, (\sigma_2^2 + \sigma_1^2\Delta_i)\mathbf{I}_{r_i})$. In essence, the $n-p$ dimensional vector $\mathbf{H}'\mathbf{Y}$ can be partitioned into d independent vectors, namely, $\mathbf{P_i'}\mathbf{H}'\mathbf{Y}$, $i = 1, ..., d$, where each sub-vector is of length r_i.

The corresponding quadratic forms associated with the independent pieces of information are $\mathbf{Y}'(\mathbf{HP_iP_i'H}')\mathbf{Y} = Q_i \sim (\sigma_2^2 + \sigma_1^2\Delta_i)\chi^2(r_i)$, $i = 1, ..., d$. By construction, the quadratic forms $Q_1, ..., Q_d$ are independent. In addition, they are a set of minimal sufficient statistics associated with the reduced linear model void of the fixed effect. The REML estimators of σ_1^2 and σ_2^2 may be obtained by maximizing the likelihood function of $Q_1, ..., Q_d$. The REML estimator of ρ is simply the corresponding function of the REML estimators of

σ_1^2 and σ_2^2. Alternatively, if ρ is the parameter under study one may rewrite the distribution of Q_i in terms of ρ and the nuisance parameter σ_2^2, to obtain $Q_i \sim \sigma_2^2(1+\Delta_i\rho/(1-\rho))\chi^2(r_i)$, $i = 1, ..., d$. In either case, only for the simplest of models will the REML estimators result in closed-form expressions.

The log-likelihood function of ρ and σ_2^2 based on the quadratic forms $Q_1, ..., Q_d$, which is denoted by $logL(\rho, \sigma_2^2; Q_1, ..., Q_d)$, is

$$\frac{1}{2}log\left(\frac{1-\rho}{\sigma_2^2}\right)\sum_{i=1}^{d} r_i - \frac{1-\rho}{2\sigma_2^2}\sum_{i=1}^{d}\frac{Q_i}{1+\rho(\Delta_i-1)} - \frac{1}{2}\sum_{i=1}^{d} r_i log(1+\rho(\Delta_i-1)) \tag{2}$$

plus other terms not involving ρ and σ_2^2. The REML estimator for ρ may be written as

$$\widehat{\rho} \;=\; \frac{\sum_{i=1}^{d}\left(\sum_{j=1}^{d}\frac{r_j(\Delta_j-\Delta_i)}{(1+\rho(\Delta_i-1))^2(1+\rho(\Delta_j-1))^2}\right)Q_i}{\sum_{i=1}^{d}\left(\sum_{j=1}^{d}\frac{r_j(\Delta_j-1)(\Delta_i-\Delta_j)}{(1+\rho(\Delta_i-1))^2(1+\rho(\Delta_j-1))^2}\right)Q_i} \tag{3}$$

where ρ is represents an initial value and an iterative method is employed until the procedure converges to a solution. Alternatively, one may use the restricted log-likelihood function and employ iterative procedures such as Newton-Raphson, expectation-maximization, method of scoring, or other algorithms to compute the REML estimator. In any case, equation (3) suggests that the REML estimator of ρ is a ratio of a weighted linear combination of the quadratic forms. The weights themselves depend on the eigenvalues, the replication of the eigenvalues, and ρ. It is interesting to note that for the simple case $d = 2$,

$$\widehat{\rho} \;=\; \frac{r_1Q_2 - r_2Q_1}{(\Delta_2-1)r_2Q_1 - (\Delta_1-1)r_1Q_2}. \tag{4}$$

By definition, the REML estimator of a variance component is confined to the corresponding parameter space. Thus, in those instances where the right side of (4) is less than zero, $\widehat{\rho}$ is set equal to zero.

It is interesting to note that in the balanced one-way random effects model, $d = 2$, Q_1 is the sum of squares within groups (error), and Q_2 is the sum of squares between groups (model). Furthermore, $\Delta_1 = 0$, r_1 is equal to the degrees of freedom within groups (error), Δ_2 is equal to the number observations per group, and r_2 is equal to the degrees of freedom between groups (model).

The asymptotic distribution of the REML estimator of ρ can be determined using the standard regularity conditions. Fisher's information matrix may be obtained from $logL(\rho, \sigma_2^2; Q_1, ..., Q_d)$. The asymptotic variance of the REML estimator of ρ is obtained by inverting Fisher's information matrix. It can be shown that the asymptotic variance of

$\widehat{\rho}$ is

$$Var(\widehat{\rho}) = \frac{2}{\sum_{i=1}^{d} r_i\left(\frac{\Delta_i-1}{1+\rho(\Delta_i-1)}\right)^2 - \frac{\left(\sum_{i=1}^{d}\frac{r_i(\Delta_i-1)}{1+\rho(\Delta_i-1)}\right)^2}{\sum_{i=1}^{d} r_i}}. \quad (5)$$

2 Closed-form Approximation to the REML Estimator of ρ

In general, a closed-form expression for the REML estimator of ρ does not exist. For $d = 2$, however, we know that a simple analytic expression for $\widehat{\rho}$ is given by (4). For more complicated cases, that is, when $d > 2$, it may be advantageous to compress that quadratic forms $Q_1, ..., Q_d$ into two quadratic forms in order to achieve an estimator having a form similar to that given in (4). This compressing of information requires one to approximate the distribution of the resulting two quadratic forms.

Recall that

$$Q_i \sim \frac{\sigma_2^2}{1-\rho}(1+\rho(\Delta_i-1))\chi^2(r_i). \quad (6)$$

A method which yields a closed-form approximation to the REML estimator of ρ is to partition $Q_1, ..., Q_d$ into two sums, namely, $\sum_{i=1}^{k} Q_i$ and $\sum_{i=k+1}^{d} Q_i$. Using Satterthwaite's (1946) method, it follows that

$$\sum_{i=1}^{k} Q_i \overset{approx}{\sim} \frac{\sigma_2^2}{1-\rho}(1+\rho(\overline{\Delta}_B-1))\chi^2(\sum_{i=1}^{k} r_i) \quad (7)$$

$$\sum_{i=k+1}^{d} Q_i \overset{approx}{\sim} \frac{\sigma_2^2}{1-\rho}(1+\rho(\overline{\Delta}_T-1))\chi^2(\sum_{i=k+1}^{d} r_i) \quad (8)$$

where

$$\overline{\Delta}_B = \frac{\sum_{i=1}^{k} r_i\Delta_i}{\sum_{i=1}^{k} r_i} \text{ and } \overline{\Delta}_T = \frac{\sum_{i=k+1}^{d} r_i\Delta_i}{\sum_{i=k+1}^{d} r_i}. \quad (9)$$

The closed-form approximation to the REML estimator of ρ indexed by k is

$$\widehat{\rho}_k = \frac{\sum_{i=1}^{k} r_i \sum_{i=k+1}^{d} Q_i - \sum_{i=k+1}^{d} r_i \sum_{i=1}^{k} Q_i}{(\overline{\Delta}_T-1)\sum_{i=k+1}^{d} r_i \sum_{i=1}^{k} Q_i - (\overline{\Delta}_B-1)\sum_{i=1}^{k} r_i \sum_{i=k+1}^{d} Q_i} \quad (10)$$

where $\widehat{\rho}_k$ is confined to the parameter space. Note that (10) is similar in form to (4) and the two equations are identical to one another when $d = 2$. It can be shown that the asymptotic variance of $\widehat{\rho}_k$ is

$$Var(\widehat{\rho}_k) = \frac{2(1+\rho(\overline{\Delta}_B - 1))^2(1+\rho(\overline{\Delta}_T - 1))^2}{(\overline{\Delta_T} - \overline{\Delta}_B)^2} \times \left[\frac{\sum_{i=k+1}^{d} r_i(1+\rho(\Delta_i - 1))^2}{\left(\sum_{i=k+1}^{d} r_i\right)^2 (1+\rho(\Delta_T - 1))^2} + \frac{\sum_{i=1}^{k} r_i(1+\rho(\Delta_i - 1))^2}{\left(\sum_{i=1}^{k} r_i\right)^2 (1+\rho(\Delta_B - 1))^2} \right]. \quad (11)$$

It is important to note that the choice of partitioning $Q_1, ..., Q_d$ into two sums is not unique and is determined by selecting the value of k. The question that arises is what is the best value of k? That is, what particular value of k results in a closed-form estimator that best mimics the characteristics of the true REML estimator of ρ? The approach taken in this paper is to compare the large sample variations of the estimators $\widehat{\rho}_k$ and $\widehat{\rho}$. Specifically, let $ASE(\widehat{\rho}_k) = \sqrt{Var(\widehat{\rho}_k)}$ and $ASE(\widehat{\rho}) = \sqrt{Var(\widehat{\rho})}$ be the asymptotic standard errors of $\widehat{\rho}_k$ and $\widehat{\rho}$, respectively. For different values of k, $ASE(\widehat{\rho}_k)$ and $ASE(\widehat{\rho})$ are compared to one another in the following application.

3 Loineye Muscle Area of Yearling Bulls

Data were obtained on one hundred and seventy one yearling bulls from a Red Angus seed stock herd in Montana (Evans et al. (1995)). One of the traits of interest was the loineye (i.e., ribeye) muscle area measured in square inches. Ultrasound techniques were used to procure these measurements which were located on the dorso-ventral line between the 12^{th} and 13^{th} ribs on the left side of each animal.

The fixed effect was age of dam which had been originally recorded as belonging to one of eight categories: 2 years, 3 years, 4 years, 5-9 years, 10 years, 11 years, 12 years, and 13 or more years. Since there were only a few observations associated with dams greater or equal 10 years of age, our analysis used five categories for age of dam: 2 years, 3 years, 4 years, 5-9 years, and 10 or more years.

The mixed linear model we consider is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \mathbf{e}$, where $\mathbf{Y}$ is a 171×1 vector of loineye muscle area measurements, $\mathbf{X}$ is a 171×5 incidence matrix, $\boldsymbol{\beta}$ is a 5×1 vector of parameters associated with the five age categories of the dams, $\mathbf{Z}$ is 171×171 identity matrix, $\mathbf{u}$ is a 171×1 vector which models the (additive) genetic effect of each animal on its loineye muscle area, and $\mathbf{e}$ is a 171×1 vector which takes into account all the other influences on loineye muscle area that have not been accounted for by the age of dam and the (additive) genetic effects.

The relationship matrix $\mathbf{A}$ was determined using a recursive method given in Henderson (1976). It uses knowledge of the animal's sire, dam, and grandparents. Note that some

animals are inbred so that it is possible that $Var(u_i) > \sigma_1^2$. For instance, it turns out that $Var(u_1) = 1.03125\sigma_1^2$. The number of distinct eigenvalues of $\mathbf{H'ZAZ'H}$ is $d = 165$. Eigenvalues range in magnitude from $\Delta_1 = 0.56569$ to $\Delta_{165} = 8.65925$. Except for $\Delta_{61} = 0.67188$ having $r_{61} = 2$, all eigenvalues have a multiplicity of one.

Since $d = 165$, there are 164 ways to partition the information $Q_1, ..., Q_{165}$ into $\sum_{i=1}^{k} Q_i$ and $\sum_{i=k+1}^{165} Q_i$ in order to obtain a closed-form approximation to the REML estimator of ρ as given in (10). We select the value of k, and hence the estimator $\widehat{\rho}_k$, that results in $ASE(\widehat{\rho}_k)$ being close to $ASE(\widehat{\rho})$. Figure 1 displays the asymptotic standard errors of $\widehat{\rho}_k$ for selected values of k and the asymptotic standard error of the REML estimator of ρ. It is not surprising to see that the values of asymptotic standard error of the estimators depend on the value of ρ.

From Figure 1, we see that the closed-form estimators $\widehat{\rho}_k$ have larger asymptotic standard errors than the REML estimator as some information was lost when compressing $Q_1, ..., Q_{165}$ into $\sum_{i=1}^{k} Q_i$ and $\sum_{i=k+1}^{165} Q_i$. Furthermore, there does not appear to be a single best $\widehat{\rho}_k$ when considering the entire parameter space from 0 to 1. One can, however, provide recommendations as to which closed-form estimators perform well in the sense of comparing $ASE(\widehat{\rho}_k)$ to $ASE(\widehat{\rho})$ across the parameter space. When making these comparisons, it is also important to note that dramatic differences between $ASE(\widehat{\rho}_k)$ and $ASE(\widehat{\rho})$ exist when ρ is small. Furthermore, if one uses the relative difference between the asymptotic standard errors as a measure of goodness of the closed-form estimator of ρ, then absolute differences between $ASE(\widehat{\rho}_k)$ and $ASE(\widehat{\rho})$ when ρ is close to zero are more important than absolute differences between $ASE(\widehat{\rho}_k)$ and $ASE(\widehat{\rho})$ when ρ is close to one since $ASE(\widehat{\rho})$ is an increasing function of ρ. For these reasons the authors suggest that the closed-form estimator associated with $k = 150$ is a viable candidate whose large sample performance tends to mimic those of the true REML estimator. In this example, it is interesting to note that the value of the true REML estimator is 0.10 whereas $\widehat{\rho}_{150} = 0.08$.

4 Summary

An analytic expression which approximates the REML estimator of heritability provides a useful alternative to a full implementation of the REML iterative procedure. The properties of the closed-form estimator mimics those of the true REML estimator. Since there are many ways in which to compress the quadratic forms into two groups, there are many possible closed-formed estimators. The closed-form approximations to the REML estimator are denoted by $\widehat{\rho}_k$. The large sample criterion used in this paper to determine the optimal choice of k considers the standard error of the closed-formed estimator as compared to the standard error of the REML estimator. In this manner one may quantify how much information is lost by using a closed-form approximation. Further research is needed in order to judge the quality of the closed-formed estimators in small sample applications.

References

Burch, B. D. and Iyer, H. K. (1997), Exact confidence intervals for a variance ratio (or heritability) in a mixed linear model, *Biometrics* **53**, 176–190.

Evans, J. L., Golden, B. L., Bailey, D. R. C., Gilbert, R. P., and Green, R. D. (1995), Genetic parameter estimates of ultrasound measures of backfat thickness, loineye muscle area, and gray shading score in red angus cattle, *Proceedings, Western Section, American Society of Animal Science* **46**, 202–204.

Harris, I. R. and Burch, B. D. (2000), Pivotal estimation with applications for the intraclass correlation coefficient in the balanced one-way random effects model, *Journal of Statistical Planning and Inference* **83**, 257–276.

Harville, D. A. (1977), Maximum likelihood approaches to variance component estimation and to related problems, *Journal of the American Statistical Association* **72**, 320–340.

Henderson, C. R. (1976). A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values, *Biometrics*, **32**, 69–83.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components, *Biometrics Bulletin*, **2**, 110–114.

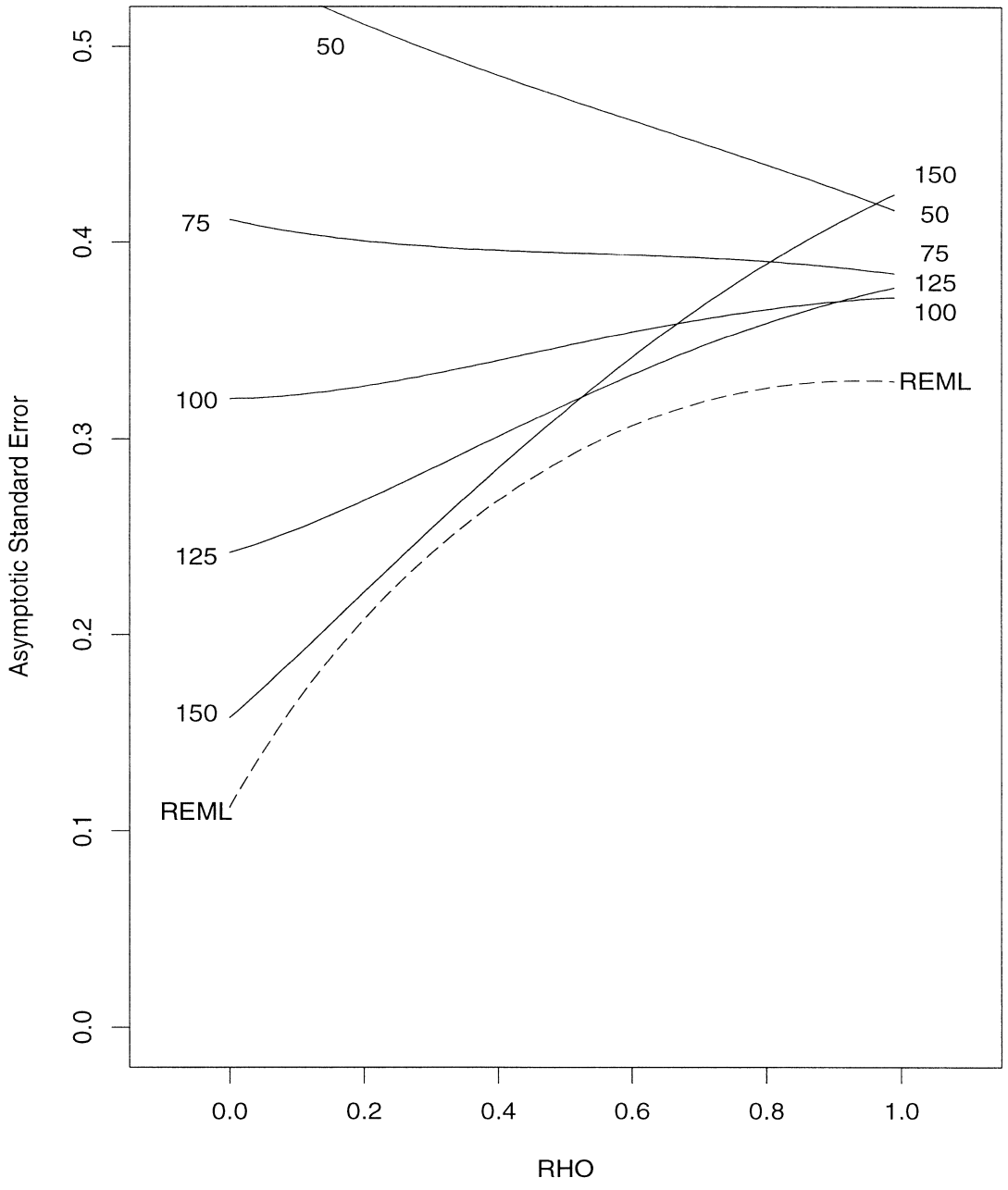

Figure 1: Asymptotic standard error of estimators of heritability