Theor Appl Genet (1983) 66:217–220

THEORETICAL AND APPLIED GENETICS

The use of the relationship matrix to account for genetic drift variance in the analysis of genetic experiments

D. A. Sorensen and B. W. Kennedy

Department of Animal and Poultry Science, University of Guelph, Guelph, Ontario N1G 2W1, Canada

Received April 28, 1982; Accepted March 20, 1983
Communicated by D. Van Vleck and A. Robertson

Summary. Selection experiments can provide information on genetic parameters such as realized heritability and response to selection. Often, due to lack of adequate replication, empirical sampling variances of estimated response cannot be computed and therefore use must be made of theoretical formulae. Most of the variance between a conceptually large number of selected lines drawn from the same base population is contributed by genetic drift, which depends on the population structure and can therefore be predicted before the experiment is carried out. The theory of variation of response to selection has been developed mainly by Hill, who produced formulae to adjust the variance of estimators to take account of genetic drift. In this paper, we draw attention to properties of the additive genetic relationship matrix that lead to well established results in population genetics theory. We show how inclusion of the additive genetic relationship matrix among the observations leads to sampling variances of estimators of genetic means that account for the variance due to genetic drift.

Key words: Response to selection – Genetic drift – Relationship matrix

Introduction

The results of selection experiments are often used to obtain estimators of realized heritabilities and related parameters, such as genetic means and response per generation. When inferences are to be made from one line to the behaviour of a conceptually large number of similarly selected lines drawn at random from the same base population, the variance of the various estimators must take account the correlated structure among the observations. Failure to do this will result in sampling variances which can be severely biased downwards, as shown by Hill (1971).

The genetic nature of the correlated error structure is due to genetic drift. In random mating populations, the process of genetic drift is well understood. Gene frequency changes due to drift in different generations are independent, but cumulative drift in a particular generation is the result of the sum of random deviations in all previous generations. Hence, the variance of the genetic mean increases each generation and means of different generations become correlated.

Variation between means of directionally selected lines is less well understood; the problem has been recently discussed by Hill (1977). Relative to unselected lines with the same effective number of males and females used as parents each generation, selection leads to the following phenomena. Firstly, selected individuals tend to be genetically more alike than randomly chosen ones and this effect will tend to reduce the variance of response. Secondly, the within-line genetic variance differs between lines due to finite population size (Avery and Hill 1977) and results in real differences in response in different lines. This effect will tend to increase the variance between lines. Finally, directional selection causes changes in gene frequency and negative covariances of gene frequencies in gametes, i.e. negative linkage disequilibrium. The latter leads to a reduction in the additive genetic variance within lines (Bulmer 1971) and this will decrease variance between lines. Changes in gene frequency can have an effect on drift variance in either direction, depending upon the initial distribution of gene effects and frequencies. All these phenomena have opposing effects on the variance of selection response and a simple operational compromise is to assume that they cancel each other out approximately. Thus, Robertson (1977) found that, up to inbreeding coefficients of 0.50, the simple formula for the drift variance in unselected lines gave a reasonable fit to his Monte Carlo selected lines under a genetic model that assumes an infinite number of loci.

In any one replicate, the effect of drift on the genetic mean cannot be predicted but the magnitude of variance between lines due to genetic drift can be quantified before the experiment is carried out from knowledge of the population structure. Hill (1971) used essentially this a priori approach to develop formulae to adjust the variance of various genetic parameter estimators. Alternatively, if the selection experiment has been conducted and pedigrees have been kept, the correlated structure among the records can be adequately described by means of the numerator relationship matrix. The purpose of this paper is to show how inclusion of the matrix of additive genetic relationships among individuals in the computation of sampling variances of estimates of genetic means accounts for variance due to genetic drift.

We assume for simplicity that the only source contributing to the correlation among the observations is their additive genetic relationship and in this paper we ignore complications associated with changes of within line genetic variance due to selection, though this problem needs to be studied further.

The relationship matrix

The relationship matrix for a group of animals is defined as the matrix with the i jth off-diagonal element equal to the numerator of Wright's (1922) coefficient of relationship of the ith and jth animals and with the ith diagonal element equal to $1 + F_i$, where F_i is the coefficient of inbreeding of the ith animal.

Consider the following model:

$$y_{ij} = m_i + g_{ij} + \varepsilon_{ij} \tag{1}$$

$i=0, \ldots, t$; $j=1, \ldots, M$; $T=t\,M$, where y_{ij} is the record on the ijth individual, m_i is the mean additive genetic value of the ith generation, g_{ij} is the additive genetic value of the jth individual in the ith generation and ε_{ij} is its environmental effect. In matrix notation, we write (1) as

$$\mathbf{y} = \mathbf{X}\,\mathbf{b} + \mathbf{Z}\,\mathbf{u} + \boldsymbol{\varepsilon} \tag{2}$$

where $\mathbf{y}$ is the vector of T observations, $\mathbf{b}$ is the vector of generation effects, $\mathbf{u}$ is the vector of random additive genetic values, $\boldsymbol{\varepsilon}$ is the vector of random environmental values and $\mathbf{X}$ and $\mathbf{Z}$ are incidence matrices. For purposes of computation we treat $\mathbf{b}$ as fixed and assume that $E(\mathbf{y}) = \mathbf{X}\,\mathbf{b}$ and

$$\mathrm{Var}(\mathbf{y}) \equiv \mathbf{V} = \mathbf{Z}\mathbf{A}\mathbf{Z}'\,\sigma_G^2 + \mathbf{I}\,\sigma_\varepsilon^2 \tag{3}$$

where σ_G^2 and σ_ε^2 are the additive genetic and environmental variances respectively in the base population, $\mathbf{I}$ is the identity matrix and $\mathbf{A}$ is the additive genetic relationship matrix each of order T×T. With one record per individual, $\mathbf{Z}$ in (2) is equal to $\mathbf{I}$.

The T observations have been generated as follows. At generation 0, M/2 males and M/2 females, mutually unrelated, are sampled from a base population in Hardy-Weinberg and linkage equilibrium. From these M individuals, Nm males and Nf females (N = Nm + Nf) are randomly chosen and mated to produce M/2 males and M/2 females of generation 1. This procedure is followed for t generations so that a total of T observations are available.

If the correlated structure among the observations is not taken into account, and the vector of random effects, $\mathbf{u}$, in (2) is ignored, such that $\mathrm{Var}(\mathbf{y})$ is incorrectly assumed to be $\mathbf{I}\,\sigma^2$, where σ^2 is the phenotypic variance, the usual least squares estimator of $\mathbf{b}$ in (2) is

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \bar{\mathbf{y}} \tag{4}$$

where $\bar{\mathbf{y}}$ is the vector of order t + 1 of raw generation means. The ith element in $\bar{\mathbf{y}}$ is: $\bar{y}_{i\cdot} = \sum_{j=1}^{M} y_{ij}/M$.

$\hat{\mathbf{b}}$ in (4) is an unbiased estimator of $\mathbf{b}$, and its correct sampling variance, given the assumptions of the model specified in (3), is:

$$\mathrm{Var}(\hat{\mathbf{b}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\sigma_G^2 + (\mathbf{X}'\mathbf{X})^{-1}\sigma_\varepsilon^2 \tag{5}$$

$$= \begin{pmatrix} \bar{a}_{00} & \bar{a}_{01} \ldots & \bar{a}_{0t} \\ \bar{a}_{10} & \bar{a}_{11} \ldots & \bar{a}_{1t} \\ \vdots & \vdots & \vdots \\ \bar{a}_{t0} & \bar{a}_{t1} \ldots & \bar{a}_{tt} \end{pmatrix} \sigma_G^2 + \mathbf{I}\,\sigma_\varepsilon^2/M \tag{6}$$

where $\bar{a}_{ij}$ is the average additive relationship between the M individuals of generation i and the M of generation j ($i = 0, \ldots, t$; $j = 0, \ldots, t$) relative to the base population. Hence, at generation zero, the variance of the least squares estimator of m_0 in (1) is

$$\mathrm{Var}(\bar{y}_0) = \bar{a}_{00}\sigma_G^2 + \sigma_\varepsilon^2/M = h^2\sigma^2/M + \sigma^2(1-h^2)/M = \sigma^2/M$$

where h^2 and σ^2 are the heritability and phenotypic variance respectively. Also, when $h^2 = 0$, (5) reduces to $(\mathbf{X}'\mathbf{X})^{-1}\sigma_\varepsilon^2$, the usual variance of the least squares estimator.

We now define a matrix $\mathbf{P}$ of order M×M which describes the flow of genes from individuals of generation i to those of generation i + 1. Matrices of this type were introduced into population genetics by Hill (1972, 1974). We can partition $\mathbf{P}$ into blocks which correspond to the pathways of genes:

$$\left(\begin{array}{c|c} \text{males (i) to} & \text{males (i) to} \\ \text{males (i + 1)} & \text{females (i + 1)} \\ \hline \text{females (i) to} & \text{females (i) to} \\ \text{males (i + 1)} & \text{females (i + 1)} \end{array}\right)$$

where rows correspond to parents in generation i and columns to progeny in generation i + 1. The elements of **P** are defined as the proportion of genes in animals at time i + 1 coming from animals at time i. Because each individual receives one gamete from each parent, the elements of each column of **P** add to one. In our model of non-overlapping generations, elements of **P** are either 0 or 1/2.

Let $\mathbf{P}_i$ denote the matrix relating individuals of generation i to those of generation i + 1. Then the matrix of additive genetic relationships among individuals of generation i and j + 1, $\mathbf{A}_{i(j+1)}$, can be represented by (Thompson 1977)

$$\mathbf{A}_{i(j+1)} = \mathbf{A}_{ij}\,\mathbf{P}_i \quad (i \leqq j)\,.$$

It can be verified that the sum of the M^2 elements in $\mathbf{A}_{i(j+1)}$ is equal to the sum of M^2 elements in $\mathbf{A}_{ii}$ $(0 \leqq j \leqq t;\ i \leqq j)$. Hence it follows that the variance of the least squares estimator of generation means is:

$$\mathrm{Var}\,(\hat{\mathbf{b}}) = \begin{pmatrix} \bar{a}_0 & \bar{a}_0 \ldots \bar{a}_0 \\ \bar{a}_0 & \bar{a}_1 \ldots \bar{a}_1 \\ \vdots & \vdots \qquad \vdots \\ \bar{a}_0 & \bar{a}_1 \ldots \bar{a}_t \end{pmatrix} \sigma^2_G + \mathbf{I}\,\sigma^2_\varepsilon/M \tag{7}$$

where $\bar{a}_i$ is the average relationship among the M individuals of generation i including relationship to self $(i = 0, \ldots, t)$.

Consider the submatrix of the additive relationship matrix **A**, corresponding to the relationship among the M individuals of a particular generation, i, say. We denoted this by $\mathbf{A}_{ii}$. We partition $\mathbf{A}_{ii}$ into 4 blocks, corresponding to the relationship between males and males, (m m), males and females (m f) and so on. Averaging within blocks, we can write (dropping the subscript i):

$$\begin{pmatrix} \bar{a}_{mm} & \bar{a}_{mf} \\ \bar{a}_{fm} & \bar{a}_{ff} \end{pmatrix}_{2\times 2}, \text{ with } \bar{a}_{mm} = \bar{a}_{ff} \text{ and } \bar{a}_{mf} = \bar{a}_{fm}\,.$$

For any generation, i, say

$$\bar{a}_i = \tfrac{1}{2}\,(\bar{a}_{mm} + \bar{a}_{mf})\,; \quad (i = 0, \ldots, t) \tag{8}$$

In the block associated with $\bar{a}_{mm}$, the k^{th} diagonal element is $1 + F_k$ whereas the corresponding element in the block associated with $\bar{a}_{mf}$ is, for the family structure we have assumed, the additive relationship between full-sibs in an inbred population which has expectation $1/2\,(1 + \bar{F}_{i-1} + 2\,\bar{F}_i)$. Hence we can write,

$$\bar{a}_{mm} = \bar{a}_{mf} + (1 - \bar{F}_{i-1})/M$$

where $\bar{F}_i$ is the average inbreeding coefficient in the i^{th} generation. Substituting in (8):

$$\bar{a}_i = \bar{a}_{mf} + (1 - \bar{F}_{i-1})/2\,M\,. \tag{9}$$

Since the Nm males and Nf females are chosen at random, the average relationship between them is equal to $\bar{a}_{mf}$. Further since the average relationship between parents is equal to twice the inbreeding coefficient of their offspring, from (7), the variance of the least squares estimator of the i^{th} element (generation mean) in $\hat{\mathbf{b}}$ is:

$$\mathrm{Var}\,(\bar{y}_{i.}) = \mathrm{Cov}\,(\bar{y}_{i.}, \bar{y}_{j.}) = \bar{a}_i\,\sigma^2_G + \sigma^2_\varepsilon/M \quad (i < j) \tag{10}$$

$$= 2\,\bar{F}_{i+1}\,\sigma^2_G + (1/2\,(1 - \bar{F}_{i-1})\,\sigma^2_G + \sigma^2_\varepsilon)/M \tag{11}$$

which reduces to σ^2_ε/M with $h^2 = 0$.

The first term in (11) is the variance due to drift, which accumulates each generation, and the second term is the error variance due to sampling a finite number (M) of offspring from N parents and it also includes the environmental variance. The error variance does not accumulate.

The term in $(1 - \bar{F}_{i-1})$ reflects the decline in within line genetic variance as inbreeding accumulates each generation. Expression (11) is equivalent to Hill's (1977) expression (5) though he chose to ignore the decline in within line genetic variance because it is small relative to changes due to drift. If an infinite number of progeny are produced expression (11) reduces to $2\,\bar{F}\ i+1\ \sigma^2_G$, which is the standard formula for the variance of the mean breeding value of N parents when the population mean is known (Falconer 1981).

Conclusions

The phenomenon of genetic drift is simply a genetic interpretation of the correlated structure among the observations and this is adequately taken into consideration by means of the relationship matrix. Non-genetic sources of correlation in the data can easily be allowed for by appropriate definition of the variance-covariance matrix of the environmental effects in the model.

The use of the relationship matrix, being essentially a retrospective approach accounts for the reduction in effective population size due to selection (Robertson 1961) with associated increase in drift variance. To arrive at the expression for the variance due to drift in (11) we had to assume a given family structure. It should be clear, however, that expression (5) is perfectly general and can be used regardless of the distribution of family size.

Although the relationship matrix can be useful in estimating the variance of the least squares estimator of genetic means, the least squares estimator, although computationally simple, can be biased under certain circumstances. A general discussion of problems associated with the estimation of genetic trend shall be the subject of a future paper.

Acknowledgements. We wish to thank Dr. W. G. Hill for advice and the Natural Sciences and Engineering Research Council of Canada for financial support.

References

Avery PJ, Hill WG (1977) Variability in genetic parameters among small populations. Genet Res 29:198–213

Bulmer MG (1971) The effect of selection on genetic variability. Am Nat 105:201–211

Falconer DS (1981) Introduction to quantitative genetics. Longmans, New York

Hill WG (1971) Design and efficiency of selection experiments for estimating genetic parameters. Biometrics 27:298–311

Hill WG (1972) Effective size of populations with overlapping generations. Theor Popul Biol 3:278–289

Hill WG (1974) Prediction and evaluation of response to selection with overlapping generations. Anim Prod 18:117–139

Hill WG (1977) Variation in response to selection. In: Int Conf Quant Gen. Iowa State University Press, Ames, pp 21–30

Robertson A (1961) Inbreeding in artificial selection programmes. Genet Res 2:189–194

Robertson A (1977) Artificial selection with a large number of linked loci. In: Int Conf Quant Gen. Iowa State University Press, Ames, pp 102–105

Thompson R (1977) The estimation of heritability with unbalanced data. 2. Data available on more than two generations. Biometrics 33:497–504

Wright S (1922) Coefficients of inbreeding and relationship. Am Nat 56:330–338