Genet Sel Evol (1995) 27, 335-345

Original article

Bias in multiple genetic correlation from half-sib designs

PM Visscher

Roslin Institute (Edinburgh), Roslin, Midlothian, EH25 9PS, UK

(Received 6 June 1994; accepted 15 March 1995)

Summary – Mean squares and mean crossproducts between and within sires were simulated to investigate the bias in genetic R^2 (defined as the square of the multiple correlation between a single trait and $(q-1)$ other traits calculated from an estimate of the genetic covariance matrix) from balanced half-sib designs. Approximate prediction equations for this bias were derived when the population correlation was zero. In that case the bias is, approximately, inversely proportional to the degrees of freedom for estimating sire components and the reliabilities of the (implicit) progeny test, and proportional to $(q-1)$. Using a genetic multiple regression based on a large number of traits and/or a small number of sires could lead to loss in response to selection relative to using a regression based on the true population parameters.

genetic regression / bias / multiple correlation / REML / half-sib design

Résumé – Biais de la corrélation génétique multiple (R^2) dans un plan expérimental avec demi-frères. *Des carrés et des co-produits moyens entre pères et intra-père ont été simulés pour étudier le biais du R^2 génétique (défini comme le carré de la corrélation multiple entre un caractère et $(q-1)$ autres caractères calculée à partir d'une estimée de la matrice des covariances génétiques), dans des schémas expérimentaux équilibrés comprenant des demi-frères. Des équations approximatives de prédiction de ce biais ont été établies dans le cas d'une corrélation nulle dans la population. Dans le cas, le biais est à peu près inversement proportionnel aux degrés de liberté d'estimation des composantes paternelles, à la précision de l'épreuve de descendance (implicite), et proportionnel à $(q-1)$. Si on utilise une régression génétique multiple basée sur un grand nombre de caractères et/ou un petit nombre de pères, on s'expose à une perte de réponse à la sélection par rapport à l'utilisation d'une régression basée sur les vrais paramètres de la population.*

régression génétique / biais / corrélation multiple / REML / schéma demi-frères

INTRODUCTION

In animal breeding, some traits are difficult or impossible to measure on animals that we want to select. For example, traits may be sex-limited (*eg*, litter size in pigs, milk production in dairy cattle), or animals may be too old by the time the trait is expressed (*eg*, herdlife in dairy cattle).

One way to predict these traits of interest is by using a regression on traits that are easier to measure. Such traits may be physiological predictors, genetic markers, or general traits which are cheaper or easier to measure (*eg*, type traits in dairy cattle to predict herdlife). In practice, the regression will most likely be a genetic regression, *ie* predicting the estimated breeding value (EBV) of the trait of interest from EBV of other traits. The use of multiple genetic markers to predict some quantitative trait is also a form of multiple (genetic) regression.

One parameter which summarizes the precision of the genetic regression is the multiple genetic correlation ρ_g, or rather its square, ρ_g^2, which is more convenient to use. We define R_g^2 as an estimate of ρ_g^2. However, it is well known that the estimate (R^2) of the squared multiple correlation coefficient (ρ^2) from phenotypic regression is biased (Fisher, 1924). The aim of this study is to investigate the behaviour of R_g^2 for balanced half-sib designs. For given population structures, intensity of selection, and relative economic values, ρ_g^2 determines the responses to selection (*eg*, Sales and Hill, 1976). Hence by investigation of the behaviour of R_g^2 we can also give examples of the consequences for selection response.

METHODS

Throughout we assume multivariate normality of observations.

Phenotypic regression

We have q traits in total, N observations, and predict the kth trait from $(q-1)$ other traits. Fisher (1924, 1928) and Wishart (1931) showed that:

$$\mathrm{E}(R^2) = 1 - \{(N-q)/(N-1)\}\ \{1-\rho^2\}\ \{\mathrm{F}[1,1,\frac{1}{2}(N-1)+1,\rho^2]\} \qquad [1]$$

where

ρ^2 = square of population multiple correlation,

F = a hypergeometric function,

$\mathrm{F}[a,b,c,x] = 1 + (ab/c)(x/1!) + \{a(a+1)b(b+1)/(c(c+1))\}\{x^2/2!\} + \ldots$

and

$$\mathrm{Var}(R^2) = \{(N-q)(N-q+2)\}/\{(N-1)(N+1)\}\ (1-\rho^2)^2\mathrm{F}[2,2,\frac{1}{2}(N-1)+2,\rho^2]$$
$$-\{(N-q)^2/(N-1)^2\}\ (1-\rho^2)^2\mathrm{F}^2[1,1,\frac{1}{2}(N-1)+1,\rho^2] \qquad [2]$$

For $\rho^2 = 0$,

$$\mathrm{E}(R^2) = (q-1)/(N-1) \text{ and} \qquad [3]$$

$$\mathrm{Var}(R^2) = \{2(q-1)(N-q)\}/\{(N-q)^2(N+1)\} \quad [4]$$

If the population correlation is not zero, approximations to the mean and variance of R^2 are

$$\mathrm{E}(R^2) \approx 1 - \{(N-q)/(N-1)\}\ \{1-\rho^2\} \text{ and}$$

$$\mathrm{Var}(R^2) \approx 4\rho^2\{(1-\rho^2)^2\}/N$$

Genetic regression

Simulation

Between ($\mathbf{B}$) and within ($\mathbf{W}$) sire matrices of mean squares and mean crossproducts of order q were simulated by sampling from independent Wishart distributions,

$$df_{\mathrm{w}}\mathbf{W} \sim \mathrm{Wishart}(df_{\mathrm{w}}, \mathbf{\Sigma}), \text{ and } df_{\mathrm{b}}\mathbf{B} \sim \mathrm{Wishart}(df_{\mathrm{b}}, \mathbf{\Sigma} + n\mathbf{\Psi}),$$

where n is the number of progeny per sire, s is the number of sires, $df_{\mathrm{w}} = s(n-1)$, and $df_{\mathrm{b}} = (s-1)$. $\mathbf{\Sigma}$ is the within-sire residual covariance matrix, and $\mathbf{\Psi}$ is the between-sire (genetic) covariance matrix. An estimate of the sire covariance matrix (which is one-fourth of the genetic covariance matrix) is

$$\mathbf{G} = (\mathbf{B} - \mathbf{W})/n$$

Parameter estimates were forced to be in the parameter space (*ie* genetic correlations between -1 and 1, and heritabilities between 0 and 1) by attenuating estimates. First $\mathbf{G}$ and $\mathbf{W}$ were diagonalised,

$$\mathbf{QGQ}' = \mathbf{D}, \quad \text{and} \quad \mathbf{QWQ}' = \mathbf{I}$$

The eigenvalues of $\mathbf{G}$, D_i, correspond to canonical heritabilities, $h^2_{ci} = 4D_i/(D_i+1)$. If canonical heritabilities were < 0 or > 1, between- and within-sire covariance matrices were attenuated as

$$\mathbf{W}^* = \mathbf{Q}^{-1}\mathbf{I}^*\mathbf{Q}'^{-1}, \quad \text{and} \quad \mathbf{G}^* = \mathbf{Q}^{-1}\mathbf{D}^*\mathbf{Q}'^{-1}.$$

If $h^2_{ci} < 0$ then $I^*_i = \{df_{\mathrm{w}} + df_{\mathrm{b}}(1+nD_i)\}/\{df_{\mathrm{w}} + df_{\mathrm{b}}\}$, and $D^*_i = 0 + \delta$, with δ a small positive number (*eg*, 10^{-6}). If $h^2_{ci} > 1$, then $I^*_i = 3/4\sigma^2$, and $D^*_i = 1/4\sigma^2$, with $\sigma^2 = \{4df_{\mathrm{w}}/3 + 4df_{\mathrm{b}}(1+nD_i)/(n+3)\}/\{df_{\mathrm{w}} + df_{\mathrm{b}}\}$. These modified variances were derived by assuming $h^2_{ci} = 0$ (or $h^2_{ci} = 1$) and re-estimating the variances from the mean squares (analogous to Thompson, 1962). This restriction procedure is similar to REML algorithms which force the estimates to be in the parameter space (*eg*, Calvin, 1993). The main reason for choosing the described restriction procedure was to reduce the amount of computing.

Without loss of generality, assume that we wish to predict trait 1 from the other $(q-1)$ traits (the predictors) using estimates of the genetic and residual covariances matrices, ρ^2_{g} is defined as

$$\rho^2_{\mathrm{g}} = (\mathbf{\Psi}'_{1\mathrm{s}}\mathbf{\Psi}^{-1}_{\mathrm{s}}\mathbf{\Psi}_{1\mathrm{s}})/\psi_{11}$$

with $\mathbf{\Psi}_{1s}$ = vector of sire covariances between trait 1 and $(q-1)$ other traits, with element Ψ_{1si} $(i = 1, \ldots, q-1)$

$\mathbf{\Psi}_s$ = sire covariance matrix for the $(q-1)$ predictors, with elements Ψ_{sij} $(i = 1, \ldots, q-1; j = 1, \ldots, q-1)$

ψ_{11} = sire variance of the trait of interest

Similarly, the estimate of ρ_g^2 is defined as

$$R_g^2 = (\mathbf{g}'_{1s}\mathbf{G}_s^{-1}\mathbf{g}_{1s})/g_{11},$$

with g_{11} the estimated sire variance of trait 1, $\mathbf{g}_{1s}$ the estimated sire covariance of trait 1 with the other traits, and $\mathbf{G}_s$ the estimated sire covariance matrix among the $(q-1)$ other traits. Both ρ_g^2 and R_g^2 are independent of whether the (estimated) sire (co)variances or the (estimated) genetic (co)variances are used.

For each set of parameters, simulation was stopped when the standard error of the mean R_g^2 was less than 0.005 (corresponding to a standard error of less than 0.5% in the tables).

Prediction

As in Sales and Hill (1976), we use a Taylor series about the true parameters to approximate the mean of R_g^2. R_g^2 is a function of $u = q(q+1)/2$ parameters,

$$\mathrm{E}(R_g^2) = R_g^2\Big|_{G=\Psi} + \sum_{i=1}^{q}\sum_{j=i}^{q}(G_{ij} - \Psi_{ij})\frac{\delta R^2}{\delta G_{ij}}\Big|_{G=\Psi} + 1/2\sum_{ij=1}^{u}\sum_{kl=1}^{u}(G_{ij} - \Psi_{ij})(G_{kl} - \Psi_{kl})\frac{\delta^2 R^2}{\delta G_{ij}\delta G_{kl}}\Big|_{G=\Psi} + \ldots \quad [5]$$

Assuming $\mathrm{E}(G_{ij}) = \Psi_{ij}$ gives,

$$\mathrm{E}(R_g^2) = \rho_g^2 + 1/2\sum_{i=1}^{q} v(G_{ii})\frac{\delta^2 R^2}{\delta G_{ii}^2}\Big|_{G=\Psi} + 1/2\sum_{i=1}^{q-1} v(g_{1si})\frac{\delta^2 R^2}{\delta g_{1si}^2}\Big|_{G=\Psi} + 1/2\sum_{i=1}^{q-1}\sum_{j=i+1}^{q-1} v(G_{sij})\frac{\delta^2 R^2}{\delta G_{sij}^2}\Big|_{G=\Psi} + 1/2\sum\sum \mathrm{cov}(G_{ij}, G_{kl})\frac{\delta^2 R^2}{\delta G_{ij}\delta G_{kl}}\Big|_{G=\Psi} + \ldots$$

For the special case of $\rho_g^2 = 0$ (and $\mathbf{\Psi}_{1s} = 0$), and assuming that $\mathbf{\Psi}$ is diagonal,

$$\mathrm{E}(R_g^2) \approx \sum_{i=1}^{q-1} v(G_{sii})(\Psi^{sii}/\psi_{11}) = \sum_{i=1}^{q-1} v(G_{sii})/(\Psi_{sii}\psi_{11})$$

with Ψ^{kk} element $(\Psi^{-1})_{kk}$. If we further assume that all $(q-1)$ predictors have equal genetic variance, *ie* $v(G_{sii}) = v(G_{sjj})$, then $\mathrm{E}(R_g^2) \approx (q-1)v(G_{skk})/(\Psi_{kk}\psi_{11})$. From multivariate theory, the variance of the covariance from a Wishart distribution is known (*eg*, Anderson, 1958, p 161). If $(df)M \sim \mathrm{Wishart}(df, \mathbf{\Sigma})$, then $v(M_{ij}) = (\sigma_{ii}\sigma_{jj} + \sigma_{ij}^2)/df$, with $\mathrm{E}(M_{ij}) = \sigma_{ij}$. Hence,

$$E(R_g^2) \approx (q-1)\{[v(B_{skk})+v(W_{skk})]/n^2\}/(\Psi_{skk}\psi_{11}) = (q-1)[1/\{(s-1)REL_1REL_p\} + \{(1-REL_1)(1-REL_p)\}/\{s(n-1)REL_pREL_1\}] \quad [6]$$

with $REL_j = n/(n+\lambda)$, $\lambda = (4-h_j^2)/h_j^2$, and REL_p the 'reliability' pertaining to the $(q-1)$ predictors. (The definition of REL_j is a standard expression of the reliability of a progeny test with n progeny and heritability h_j^2.) If $s(n-1)$ is large, then the simplest approximation to R_g^2 is

$$E(R_g^2) \approx (q-1)/\{(s-1)REL_1REL_p\} \quad [7]$$

These approximations are appealing because of their similarity to [3]. (*NB*: [6] and [7] reduce to $(q-1)/(s-1)$ for large n.) Equation [7] indicates that the expected value of the estimate of ρ_g^2 is approximately the number of variates used in the genetic regression divided by the 'effective number of sires'.

In some cases, for example when we deal with genetic markers, the heritabilities of the $(q-1)$ predictors, and their correlations with each other, may be known *a priori*. If the covariances among the predictors are zero, and their heritabilities are equal, then, after some algebra,

$$E(R_g^2) \approx (q-1)/\{(s-1)REL_1REL_p\} + \rho_g^2[1+1/(s-1)+2/\{(s-1)(REL_1^2) - 2/\{(s-1)(REL_1)\} - 2/\{(s-1)(REL_p)\}] \quad [8]$$

Equation [8] suggests an adjusted estimate of ρ_g^2,

$$R_g^{2*} = [R_g^2 - (q-1)/\{(s-1)REL_1REL_p\}]/[(s-1)+2/\{(s-1)(REL_1^2)\} - 2/\{(s-1)(REL_1)\} - 2/\{(s-1)(REL_p)\}] \quad [9]$$

RESULTS

Examples for phenotypic R^2

In table I, the exact mean and standard deviation of R^2 are given for various combinations of ρ^2, q and N (using results from Wishart (1931)). As was shown in the previous section, these values correspond to the limiting case of very large progeny group sizes in half-sib designs. For most combinations the bias in R^2 is small, although for relatively few observations ($N = 100$ and $N = 200$) and a large number of traits ($q = 10$ and $q = 20$), the bias and standard deviation of R^2 can be large. For example, when $\rho^2 = 0$ and $q = 20$, the mean and standard deviation of R^2 ($\times 100$) for $N = 100$ are 19.2 and 5.5 respectively (table II). Even for $N = 400$, the mean R^2 is nearly 0.05.

Examples for R_g^2 *when* $\rho_g^2 = 0$

In table II, simulation results, and their predictions, are shown for various combinations of s, q, and n. The predictions were made according to [6], using population

Table I. Exact mean (first rows) and standard deviation (second rows) of phenotypic R^2 ($\times 100$) for different number of variates (q) and observations (N), and population ρ^2 using equations [1] and [2].

q	$\rho^2 = 0$			$\rho^2 = 50\%$			$\rho^2 = 75\%$		
	N = *100*	N = *200*	N = *400*	N = *100*	N = *200*	N = *400*	N = *100*	N = *200*	N = *400*
2	1.0	0.5	0.3	50.0	50.0	50.0	74.9	74.9	75.0
	1.4	0.7	0.4	7.1	5.0	3.5	4.4	3.1	2.2
5	4.0	2.0	1.0	51.5	50.8	50.4	75.6	75.3	75.2
	2.8	1.4	0.7	7.0	5.0	3.5	4.3	3.1	2.2
10	9.1	4.5	2.3	54.1	52.0	51.0	76.9	75.9	75.5
	4.0	2.1	1.0	6.8	4.9	3.5	4.2	3.0	2.1
20	19.2	9.5	4.8	59.2	54.5	52.3	79.5	77.2	76.1
	5.5	2.9	1.5	6.4	4.8	3.5	3.9	2.9	2.1

Table II. Mean R_g^2 ($\times 100$) from simulation (first rows) and prediction (second rows) for different number of sires (s), progeny per sire (n), and number of variates (q), for equal heritabilities ($h^2 = 0.25$) and $\rho_g^2 = 0$.

q	n = *25*					n = *50*				
	S = *100*	S = *200*	S = *400*	S = *800*	S = *1 600*	S = *100*	S = *200*	S = *400*	S = *800*	S = *1 600*
2	2.8	0.8	0.9	0.1	0.2	1.5	0.8	0.4	0.2	0.0
	2.6	1.3	0.6	0.3	0.2	1.7	0.9	0.4	0.2	0.1
5	11.1	5.0	2.8	1.5	0.8	6.8	2.6	1.6	0.8	0.4
	10.4	5.2	2.6	1.3	0.6	6.8	3.4	1.7	0.8	0.4
10	27.9	12.5	6.4	2.7	1.3	16.7	7.3	4.1	1.4	1.0
	23.4	11.6	5.8	2.9	1.4	15.4	7.7	3.8	1.9	2.0
20	93.2	29.0	13.1	5.7	2.6	38.0	18.2	9.1	4.2	2.0
	49.4	24.6	12.3	6.1	3.1	32.5	16.2	8.1	4.0	2.0

parameters. In all cases the heritability of all traits was 0.25. In general, predictions and simulation results agreed reasonably, although for small n and s, and large q, the prediction tends to be too low. For example, for $s = 100$, $q = 20$ and $n = 25$, the average R_g^2 from simulation was 0.93, whereas the prediction was only 0.49.

Predicting herdlife from type traits

Various authors have found associations between type traits and herdlife or survival in dairy cattle (*eg*, Rogers *et al*, 1988; Brotherstone and Hill, 1991; Boldman *et al*, 1992; Short and Lawlor, 1992). Most analyses were from sire models with many type traits analysed simultaneously. A typical value for the heritability of functional herdlife (= HL = herdlife adjusted for milk production) is 0.05. Equation [7] was applied to the situation where (functional) herdlife is predicted from a range of type

traits, with h^2 of herdlife of 0.05 and h^2 of type traits of 0.30. Average predicted R_g^2 ($\times 100$) for $\rho_g^2 = 0$, $q = 20$ and $n = 50$, were 61.9, 30.8, 15.4, 7.7 and 3.8 for $s = 100$, 200, 400, 800, and 1 600, respectively.

In practice the EBV for milk yield may be combined with the EBV for herdlife (predicted from EBV of type traits) in an overall selection index. The efficiency of such an index was investigated using results from Short and Lawlor (1992). Their estimated genetic and phenotypic covariance matrices of HL and 15 type traits (hence, $q = 16$) for grade Holsteins were assumed to be the population covariance matrices. For each simulation, the estimated covariance matrices (with $s = 1\,400$ and $n = 33$) were used to create a selection index combining milk with HL. It was assumed that the h^2 for milk yield was known ($h^2 = 0.25$), and that milk yield and HL were independent (it is a separate issue what the correlation between adjusted herdlife and milk yield really is, since the adjustment is usually at the phenotypic level). Further assumptions were that the selection index was based on 50 progeny for milk yield and type traits, and that relative economic weights of milk/HL were 2:1 (in genetic standard deviation units). These results are presented in table III. The (assumed) genetic ρ_g^2 was 0.37, which follows directly from the results from Short and Lawlor (1992). The average R_g^2 from simulation was 0.81, with a proportion of 0.58 of the simulated genetic covariance matrices that were attenuated. The optimum selection index (using population covariance matrices) resulted in a correlation between index and goal (r_{IH}) of 0.813. The achieved r_{IH} was on average 0.795, and the predicted r_{IH} (assuming the estimated covariance matrices are the true ones) was 0.82 (table III). Hence, although the genetic R_g^2 was severely overestimated, the loss in response was small (0.795/0.813 = 0.978 efficiency). Ignoring type traits altogether gives an r_{IH} of 0.785. Finally, using a selection index with milk yield and HL itself results in $r_{IH} = 0.826$.

Table III. Accuracy of index selection on milk yield and traits (and herdlife) using the data of Short and Lawlor (1992).

Index	*Traits in index*	*Accuracy × 100*	*Efficiency × 100*
Optimum[a]	Milk and type	81.3	100
Achieved[b]	Milk and type	79.5	97.8
Predicted[b]	Milk and type	82.0	–
Optimum	Milk	78.5	96.6
Optimum	Milk and herdlife	82.6	101.6

[a]Assuming estimates from Short and Lawlor (1992) are true population parameters; [b]average from simulation.

DISCUSSION

For half-sib population structures, average R_g^2 obtained from simulation and from prediction equations were compared for different number of sires, number of traits, and number of progeny per sire. In general, there was good agreement, although

with a large number of traits (q) and small number of sires (s), average R_g^2 from simulation were larger than predicted. The reason for this is 2-fold. First, higher order terms from the Taylor series which are not taken into account are likely to be proportional to q^2, so that the prediction would be too low. Second, for combinations of large q and small n, the probability of non-positive-definite matrices and hence attenuation is higher (Hill and Thompson, 1978). After attenuation, the assumption of $E(\mathbf{G}) = \mathbf{\Psi}$ is not valid anymore, and the prediction will be out. For $s = 100$, $q = 20$ and $n = 25$, including higher order terms in the prediction (terms not shown) gave a predicted R_g^2 of 0.58.

In table IV simulation results are presented separately for those replicates whose estimated covariance matrices were attenuated, and for those for which no attenuation was required (*ie* $\mathbf{G} = (\mathbf{B} - \mathbf{W})/n$). For nearly all combinations of parameters, the average R_g^2 was nearly 1.0 for when covariance matrices were attenuated. This can be explained as follows: when the $(\mathbf{B} - \mathbf{W})$ is non-positive-definite, a linear combination of all traits exists with zero genetic variance, and, therefore, any single trait may be predicted from a linear combination of all other traits with an accuracy of unity. Consider the bivariate case when the linear combination $l_1y_1 + l_2y_2$ has zero variance, $\text{Var}(l_1y_1 + l_2y_2) = a + 2\text{cov} + b = 0$. Hence, $\text{cov} = -(a+b)/2$, and $r = -(a+b)/2(ab)^{1/2}$. The last term is always < -1, unless $a = b$. Hence, on the original scale, the correlation between y_1 and y_2 is < -1, which will be forced to -1, and the resulting R^2 will be 1.0. The same principle holds when for more than 2 traits, *ie* when y_2 itself is a linear combination of more than 2 traits.

Table IV. Mean R_g^2 ($\times$ 100) from simulation and prediction (equation [7] for $q = 10$ variates and different number of sires (s), and progeny per (n) for equal heritabilities ($h^2 = 0.25$) and $\rho_g^2 = 0$; results split according to proportion of covariance matrices that were attenuated.

s	n	*Attenuated (%)*	R_g^2 *(× 100)*			
			Unattenuated	*Attenuated*	*Average*	*Prediction*
100	5	100	–	100	100	100
	10	91	60	99	96	59
	15	35	49	99	66	37
	20	5	36	97	39	28
200	5	100	–	100	100	83
	10	17	41	99	51	29
	15	0	22	–	22	18
	20	0	16	–	16	14

This has implications for inferences drawn from REML estimation, because most REML algorithms in practice do require estimates to be within the parameter space. Therefore, one should be very cautious in drawing inferences about functions of parameter estimates (such as R_g^2) from large estimated covariance matrices.

Because the mean R_g^2 depends on whether covariance matrices are attenuated, a refinement of the prediction equations is to predict the proportion of estimates for which this occurs. This was beyond the scope of this study, but Hill and Thompson (1978, and references therein) addressed that issue.

Meyer and Hill (1983) found large losses in response for $s = 100$, $n = 4(8)$ and 2 or 4 traits of equal importance when estimated covariance matrices were used in a selection index. Losses in response were much smaller when 'bending' was applied to the between-sire covariance matrix.

Overestimation of the multiple correlation coefficient from a multiple regression of (estimated) breeding values on genetic marker scores has similarities with the topic addressed in this study. When estimating associations between genetic markers and quantitative traits we have to specify what kind of population the sample is from. Usually association studies are either from populations derived from crosses between divergent lines (or inbred lines) or within families in completely outbred populations. When dealing with crosses from different breeds or inbred lines, the bias in phenotypic R^2 applies since linkage disequilibrium will be across the population. For half-sib designs in outbred populations essentially the bias in the within-sire R^2 is of interest because regressions of phenotypes on markers are within families. However, these cases are extremes. In practice, we may deal with a population which was created by hybridization a number of generations ago, and in that case it would not be unreasonable to look for genetic markers that explain some of the between-sire variance. A thorough study of the bias in R^2 from using genetic markers, taking into account the discrete nature of marker scores and linkages between markers and quantitative trait loci was outside the scope of this study. Sales and Hill (1976) derived losses in response to selection when including worthless marker traits in a selection index. For marker-assisted selection in a population created by recent hybridization, re-sampling of data after choosing an initial set of markers (Lande and Thompson, 1990) should reduce the bias in R^2. However, although the individual marker effects may be estimated without bias in the subsequent sample (a result of Lande and Thompson's proposal), their combined effect, as measured by the R^2, may still be biased. This could lead to a loss in response to selection compared to using the true marker effects because information from markers will usually be combined with phenotypic information in a selection index so that an upward bias in the R^2 from markers will result in too much weight given to the marker information.

In general, obtaining unbiased estimates of ρ_g^2 is intractable, because the mean of R_g^2 depends on the unknown population parameters in a complex way (*ie* first and second derivatives of R_g^2 with respect to estimates of individual variance components in the Taylor series). In very limited cases, prior information about (co)variances can be used to adjust R_g^2. For example, if the heritabilities for all traits are known, and the $(q-1)$ predictors are known to be uncorrelated, Equation [9] can be used to adjust R_g^2. Table V shows simulation results using [9]. The adjustment works well, expect for large q and small s. The reasons for the poor performance of the adjustment for $q = 20$ and $s = 100$ are the same as before, *ie* higher order terms in the Taylor series are ignored and the probability of attenuation is higher.

Although the genetic R_g^2 for predicting herdlife may be severely overestimated, the effect on loss in response to selection seems small. This is because the relative

Table V. Mean R_g^2 ($\times$ 100) from simulation using equation [9] for different number of sires (s) and number of variates (q), for 25 progeny per sire, for equal heritabilities ($h^2 = 0.25$) and different values of ρ_g^2 ($\times$ 100).

s	$\rho_g^2 = 25$			$\rho_g^2 = 50$			$\rho_g^2 = 75$		
	q = 100	q = 400	q = 800	q = 100	q = 400	q = 800	q = 100	q = 400	q = 800
2	25.9	25.0	24.6	50.2	50.1	49.5	73.4	74.0	75.2
5	26.4	24.0	25.1	52.0	50.6	49.9	72.5	74.6	75.0
10	31.5	25.8	25.9	53.5	50.6	50.0	69.6	75.3	74.6
20	49.3	27.3	26.2	51.0	53.0	50.3	51.1	75.0	74.9

economic weight for HL was assumed to be half that of milk yield, and because the heritability of HL was small. For the example of Short and Lawlor (1992), h^2 of HL was only 0.04. Hence, even if we think we can accurately predict HL when in fact the prediction is inaccurate, response to selection is only reduced slightly because the prediction of HL gets a low weight in the overall selection index. Still, the loss in efficiency (2.2% for the example) should be compared to the maximum gain obtained by including type traits (0.813/0.785 = 3.6% extra gain in the example). Thus, only about one-third of the maximum achievable gain was obtained. Finally, it seems undesirable to include traits in the selection index for which the estimated parameters may be subject to large error.

ACKNOWLEDGMENTS

This work was funded by the Marker-Assisted Selection Consortium of the British pig industry (Cotswold Pig Development Company Ltd, JSR Farms Ltd, National Pig Development Company, Newsham Hybrid Pigs Ltd, Pig Improvement Company, and the Meat and Livestock Commission) and by MAFF, DTI, and the BBSRC. I thank M Goddard for bringing the topic to my attention when we were in Melbourne (at Carlton Place?) and for constructive comments. Thanks to R Thompson, C Haley, and B Hill for discussions and helpful comments. Special thanks to RT for deattenuating my vocabulary.

REFERENCES

Anderson TW (1958) *Introduction to Statistical Multivariate Analysis.* John Wiley & Sons, New York, USA

Boldman KG, Freeman AE, Harris BL, Kuck AL (1992) Prediction of sire transmitting abilities for herd life from transmitting abilities for linear type traits. *J Dairy Sci* 75, 552-563

Brotherstone S, Hill WG (1991) Dairy herd life in relation to linear type traits and production. 1. Phenotypic and genetic analyses in pedigree type classified herds. *Anim Prod* 53, 279-287

Calvin JA (1993) REML estimation in unbalanced multivariate variance components models using an EM algorithm. *Biometrics* 49, 691-701

Fisher RA (1924) The influence of rainfall on the yield of wheat at Rothamsted. *Phil Trans B* 213, 89-142

Fisher RA (1928) The general sampling distribution of the multiple correlation coefficient. *Proc R Soc Lond A* 121, 654-673

Hill WG, Thompson R (1978) Probabilities of non-positive definite between-group or genetic covariance matrices. *Biometrics* 34, 429-439

Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124, 743-756

Meyer K, Hill WG (1983) A note on the effects of sampling errors on the accuracy of genetic selection indices. *J Anim Breed Genet* 100, 27-32

Rogers GW, McDaniel BT, Dentine MR, Johnson LP (1988) Relationships among survival rates, predicted differences for yield, and linear types traits. *J Dairy Sci* 71, 214-222

Sales J, Hill WG (1976) Effect of sampling errors on efficiency of selection indices. 2. Use of information on associated traits for improvement of a single important trait. *Anim Prod* 23, 1-14

Short TH, Lawlor TJ (1992) Genetic parameters of conformation traits, milk yield, and herd life in Holsteins. *J Dairy Sci* 75, 1987-1998

Thompson WA (1962) The problem of negative estimates of variance components. *Ann Math Statist* 33, 273-289

Wishart J (1931) The mean and second moment coefficient of the multiple correlation coefficient, in samples from a normal population. *Biometrika* 22, 353-361