On the Estimation of Variances Within Herd-Mean Production Groups

PETER M. VISSCHER
Institute of Cell, Animal, and Population Biology[1]
University of Edinburgh
West Mains Road
Edinburgh EH9 3JT, United Kingdom

ABSTRACT

In investigating heterogeneity of variance among herd groups, herds are often grouped according to their mean production, and variances are estimated within such groups using a sire model. Stratifying herds in this way may be a form of "selection" on sire progeny groups, resulting in herd production groups with a selected sample of sires, causing a "pseudoheterogeneity" of variance to be estimated. A selection effect was not found for balanced nested and cross-classified designs. For a balanced design, the only biased estimate of the sire variance is obtained for the limited case of selection directly on progeny group means, with the absence of any fixed effects other than a mean. For unbalanced designs, the bias depends on the distribution of regressions of progeny means on herd means and is likely to be negligible in most practical analyses in which there is substantial variation within herds due to environmental effects.

(Key words: variance estimation, selection bias, heterogeneity of variance)

Abbreviation key: HYS = herd-year-season, SS = sum of squares, SSS = sire sum of squares.

INTRODUCTION

One of the assumptions usually made by users of BLUP is homogeneity of variance across fixed effect levels. There is abundant evidence, however, of heterogeneity of variance across herds or herd-year-seasons (HYS) for milk production traits in recent studies (1, 2, 5, 6, 7). Some of these authors have found a relationship between herd mean and within-herd (genetic) variance. Typically for those studies, herds were classified according to their mean (milk) production, and parameters were estimated within (and between) herd mean production groups, using a sire model.

Famula (3) argued that stratifying herds in this way can be regarded as a form of selection on sire progeny groups; herd means may be higher because of the sires represented in those herds, resulting in herd production groups with a selected sample of sires. A pseudoheterogeneity of variance could, therefore, be induced by selecting herds on their mean production (3). However, the results from the simulation study presented in Famula's paper are not clear, because the observed biases in estimated sire variances were probably not significant (standard errors were not presented, but these can be estimated from the presented ranges and the number of replicates). Furthermore, one would expect the selection effect to be symmetrical about the overall mean; i.e., a bias in estimating variances from the highest herd mean group should be similar to a bias from the lowest herd mean group. This was not observed. The aim of this study was to qualify and quantify the magnitude of this selection effect.

MATERIALS AND METHODS

For various balanced and "semibalanced" designs, the effect of selection of herd production groups on the estimates of genetic and residual variances can be quantified.

Balanced Nested Designs of Sires Within Herds

The reduction in variance between progeny groups depends on the regression of sire progeny mean on herd mean. This reduction is

Received August 27, 1990.
Accepted December 3, 1990.
[1]Formerly Genetics Department.

largest for a nested design of sires within herds, in the absence of herd effects and other fixed effects, because then selection on herd means is highly correlated with (direct) selection of sires.

Notation includes the following:

h = number of herds in selected group,
n = number of sires per herd,
p = number of progeny per sire,
Y = sire progeny mean,
H = herd mean,
subscript s = selected, and
i = mean of selected group (= selection intensity).

Normality of random effects is assumed throughout this study. Without loss of generality, let the total phenotypic variance in the base population be unity. Then

$$v(Y) = (1 - t)/p + t \quad [1]$$

$v(H) = cov(Y,H) = (1 - t)/(n \times p) + t/n$ where t is the intraclass correlation in the base population. Then

$$b_{Y,H} = 1;\ r^2_{Y,H} = 1/n.$$

Using simple linear regression

$$v(Y_s) = (1 - kr^2)v(Y), \quad [2]$$

with k being the reduction in variance for the selected group [= i(i − x)] for truncation selection, where x is the deviation of the truncation point from the mean in standard deviation units).

For a balanced design, the orthogonal sums of squares (SS) for herds, sires, and residual from the ANOVA can be equated to their expectations. It can be shown easily that the expectation for the residual variance is the error variance of the base population. The expectation of the sire sum of squares (SSS) is, on conditioning on the herd mean,

$$\begin{aligned} E(SSS|H) &= E[\{\Sigma p(Y - H)^2\}|H] \\ &= pn[E(Y^2|H) - E(H^2|H)] \\ &= pn[(1 - r^2)v(Y) + b^2H^2 - H^2] \\ &= p(n - 1)v(Y), \text{ because } r^2 = 1/n \\ &\quad \text{and } b^2 = 1. \end{aligned}$$

Therefore the SSS is not dependent on the herd value; in whatever way the herds are selected, the within-herd SSS is unbiased through that selection. The estimated variances in any selected group are, therefore, unbiased estimators of the population parameters.

Although the expectation of the sums of squares between sire progeny group means is unaffected by selection, the expectation of the variance between the unobserved sire effects is not. The reduction in genetic variance for the selected group can be predicted using the regression of sire values on herd means. It follows that

$$E[v(s)_s] = (1 - kr^2/n)v(s), \quad [3]$$

where $r^2 = p/(p + \lambda)$,
$\lambda = (1 - t)/t$ and
v(s) = sire variance in the base population.

Selection on Progeny Means

For the limited case of one sire per herd (n = 1; h = total number of sires), i.e., ignoring herds and selecting solely on progeny means, it can be shown that the expected estimated sire variance in the selected group is

$$E[\sigma^2_s] = v(s)\ [1 - k(\lambda + p)/p] \quad [4]$$

The expectations in Equations [3] and [4] are identical only in the case of no selection, i.e., k = 0, or for the trivial case of $\lambda = 0$. The term between the square brackets can become negative for $k > p/(p + \lambda)$, that is, if the repeatability of the predicted sire effect is smaller than the reduction in variance for the selected group. If the ordered progeny group means are divided into four groups by symmetric truncation about the mean, then the largest reduction in estimated sire variance is expected in the two middle groups, because the variation between progeny means is the smallest in those groups. If each group contains exactly 25% of the population, then it can be shown that the reduction in variance for the two middle groups is .95 (= k). Of course, the distribution of progeny means is symmetric so that selection of the top or the bottom groups should yield identical results. Famula (3) used

fixed truncation points to obtain four groups each containing approximately 25% of the herds.

Table 1 shows a few combinations of the number of sires in the base population (m), h^2, and p, together with predictions of estimated parameters and simulation results. Records were simulated as a sire effect plus a random error term and evaluated with an ANOVA, fitting an overall mean and a between- and within-sire term. For the examples given, the heritabilities were chosen to be large because, for low heritabilities and few daughters per sire, (highly) negative estimated sire variances were expected (for repeatability < k). The number of replicates was chosen to obtain sufficiently small standard errors of the mean estimates and varied for different sets of parameters. As expected, the simulation results agree well with the predictions. Although this model, for which the criteria on which selection took place are ignored, is unlikely to be used in practical situations, the results show that even in cases with extremely high heritabilities, negative variances may be expected.

Balanced Cross-Classified Designs

For a balanced cross-classified design, unbiased estimators of the population variances are again obtained: selection on herd means now is solely environmental, because the variation between herd means does not contain a between-sire component. Although the between-herd SS are reduced, the expectation of sire and residual SS remains the same. Famula (3) gave a generalization for the expected SSS using Henderson's method 3 in his formula 11. It is shown in the Appendix that the last two terms (the bias) in that formula reduce to zero for balanced designs. In all cases, the estimate of the residual variance is unbiased.

Semi-Balanced Nested Designs

A bias does occur, however, for unbalanced designs, because the regression of progeny means on herd means is not constant for all sires. Consider the semi-balanced case of n sires nested within herds, with p_{ij} progeny for sire j in herd i. Similarly, b_{ij} is the regression of progeny group mean j on herd mean i. Assume the distribution of progeny numbers over sires within a herd is the same for all herds; for example, all herds have $(p_1 + p_2)$ progeny records pertaining to two sires, with p_1 and p_2 constant for all herds. Let the sum of all records within a herd be m. Then

$$\begin{aligned} v(H) &= v[(\Sigma p_{ij}Y_{ij})/(\Sigma p_{ij})] \\ &= (1/\{\Sigma p_{ij}\}^2) \\ &\quad \Sigma[p_{ij}^2\,((1 - t)/p_{ij} + t)] \\ &= (1/m^2)\ \Sigma[p_{ij}\ (1 + t(p_{ij} - 1))] \end{aligned} \quad [5]$$

and

$$\mathrm{cov}(Y_{ij},H_i) = (p_{ij}/m)v(Y_{ij}). \quad [6]$$

Therefore,

$$b_{ij} = [mp_{ij}v(Y_{ij})]/[\Sigma p_{ij}(1 + t(p_{ij} - 1))]. \quad [7]$$

The SSS, on conditioning on the herd mean, is

$$\begin{aligned} E[SSS|H] &= E[\Sigma p_{ij}Y_{ij}^2|H] - E[\Sigma p_{ij}H^2|H] \\ &= \Sigma p_{ij}v(Y_{ij}) - v(H)\Sigma p_{ij}b_{ij}^2 \\ &\quad + H^2[\Sigma p_{ij}b_{ij}^2 - \Sigma p_{ij}]. \end{aligned}$$

Now the SSS can depend on the herd value H. Only for the cases of all $b_{ij} = 1$, i.e., the balanced case, or for the case of $E(H^2) = v(H)$, i.e., $E(H) = 0$, does the formula reduce to the form independent of herd means.

Averaging over all possible herd values in the selected group gives

$$\begin{aligned} E[SSS] &= \Sigma p_{ij}v(Y_{ij}) - v(H)\Sigma p_{ij} \\ &\quad + (i^2 - k)v(H)[\Sigma p_{ij}b_{ij}^2 - \Sigma p_{ij}]. \end{aligned} \quad [8]$$

The first two terms are the usual terms for this design, resulting in an unbiased estimate of the sire variance. The last term is the bias in the SS. The bias for the estimated intraclass correlation is

$$\begin{aligned} BIAS(t) = {} & \{(i^2 - k)v(H)[\Sigma p_{ij}b_{ij}^2 - \Sigma p_{ij}]\}/ \\ & \{\Sigma p_{ij} - \Sigma p_{ij}^2/\Sigma p_{ij}\}. \end{aligned}$$

TABLE 1. Observed and predicted results for selecting on progeny means.[1]

				Observed parameters						Predicted parameters using Equations [2], [4], and [3]		
m	p	h^2	Group	$v(Y_s)$		σ_s^2		v(s)		$v(Y_s)$	σ_s^2	$v(s)_s$
				$\bar{X}$	SE	$\bar{X}$	SE	$\bar{X}$	SE			
48	10	1.0	4	7.996	.064	.509	.066	10.454	.067	8.06	.56	10.54
100	25	1.0	3	1.148	.008	−1.858	.009	3.618	.019	1.15	−1.85	3.59
100	50	.50	3	.590	.007	−1.172	.007	1.973	.018	.58	−1.17	1.98

[1]Phenotypic variance simulated = 100 (units)^2; m = Number of sires in base population; p = Number of progeny per sire; groups: 4 = top 25%; 3 = second ("next") 25%; $v(Y_s)$ = Variance between progeny groups in selected group; σ_s^2 = estimated sire variance from ANOVA; $v(s)_s$ = true sire variance in selected group.

To illustrate the effect, an example is given for p_{ij} = (1,10), h^2 = .25, v(p) = 1, in the absence of true herd effects; p_{ij} = (1,10) means that each herd has 11 progeny records, one pertaining to the first sire and 10 to the second sire represented in that herd. Then, using Equations [1], [5], and [7],

$$v(H) = .1374,\ b_1 = .66 \text{ and } b_2 = 1.03.$$

Selecting the top and bottom 25% of the herds (i = 1.27, k = .77, $i^2 - k$ = .85) gives the bias in the SS of 1.458 per herd (using Equation [3]) and, hence, a bias in the estimated heritability of +.03. Selecting either of the remaining middle groups (i = .32, k = .95, $i^2 - k$ = .85) gives the bias in the heritability of −.03. These results were compared with simulation results and agreed well. Some more examples are given in Table 2.

In extreme cases, a substantial bias may occur, but for moderate heritability values and three or more sires per herd, the bias becomes very small. For the described design, the direction of the bias is determined by the sign of the factor $(i^2 - k)$. It follows that the heritability is overestimated from evaluating the top and bottom 25% herds and underestimated when selecting the "next" 25% groups, the absolute value of the bias being the same for both groups, because the quantity $|i^2 - k|$ is identical for the above groups.

For the limited case of only two sires per herd, the result becomes obvious if the covariance between the difference of the two progeny group means and the herd mean is considered. This covariance is

$$\text{cov}[(Y_1 - Y_2, H)] = t(p_1 - p_2)/(p_1 + p_2)$$

where p_1 = progeny number of sire 1 and p_2 = progeny number of sire 2. For this example,

TABLE 2. Predicted biases in heritability estimates from semibalanced nested design.[1]

Sires per herd	Progeny distribution	h^2	Estimated h^2		Bias (h^2)	
			Top	Middle	Top	Middle
2	1, 10	.25	.282	.218	.032	−.032
2	1, 10	.50	.596	.404	.096	−.096
3	1, 5, 10	.25	.259	.241	.009	−.009
3	1, 5, 10	.50	.529	.471	.029	−.029
10	1, 1, 4, 4, 5, 5, 6, 6, 9, 9	.25	.251	.249	.001	−.001
10	1, 1, 4, 4, 5, 5, 6, 6, 9, 9	.50	.504	.496	.004	−.004

[1]Top = top (or bottom) 25% herds are selected; middle = second (or third) 25% of herds.

this covariance is .051, and the regression of the progeny mean difference on the herd mean is .37 (for $p_1 > p_2$), which is the difference between the two regression coefficients. Hence, the difference between progeny groups within herds depends on the mean of that herd, although the difference between the sire values remains independent of the herd mean.

DISCUSSION

In practice, the regression of progeny mean on herd mean may well be close to zero due to HYS and other fixed effects. Therefore, the bias for the estimated parameters and the reduction in true genetic variance in the selected group will both be small. Because young sires usually are distributed over many herds, the selection effect is thought to be negligible for most practical evaluations. Famula (3) simulated 1800 HYS effects from 150 herds and 150 sire effects and randomly assigned 15,000 progeny records to (270,000) HYS by sire subclasses, resulting in an unbalanced cross-classified design. Regressions of progeny means on herd means were likely to be small, because the expected number of records per sire by herd subclass was $15{,}000/(150 \times 150) = .67$. Furthermore, the difference between those regression coefficients within any herd were probably small. Famula's results (that the higher the mean of the herd group, the lower the estimated sire variance) can, therefore, most likely be explained by sampling. In practice, there usually is substantial variation within herds due to environmental (e.g., HYS) effects; therefore, the regressions of progeny means on herd means are expected to be small. Most likely, the sire selection effect of stratifying herds on their mean production is, therefore, negligible. If high producing herds have a different sire selection strategy from low producing herds, inducing an additional covariance between sire and herd values, then heterogeneity of variance is present and will be detected by the estimation methods in use.

ACKNOWLEDGMENTS

Financial support from the Milk Marketing Boards in the UK is acknowledged. I wish to thank Robin Thompson and Bill Hill for helpful discussions and Steve Bishop for many useful comments on an earlier version of the manuscript.

REFERENCES

1 Boldman, K. G., and A. E. Freeman. 1990. Adjustment for heterogeneity of variances by herd production level in the dairy cow and sire evaluation. J. Dairy Sci. 73:503.
2 Brotherstone, S., and W. G. Hill. 1986. Heterogeneity of variance amongst herds for milk production. Anim. Prod. 42:297.
3 Famula, T. R. 1989. Detection of heterogeneous variance in herd production groups. J. Dairy Sci. 72:715.
4 Henderson, C. R. 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423.
5 Hill, W. G., M. R. Edwards, M.-K. A. Ahmed. and R. Thompson. 1983. Heritability of milk yield and composition at different levels and variability of production. Anim. Prod. 36:59.
6 Lofgren, D. L., W. E. Vinson, R. E. Pearson, and R. L. Powell. 1985. Heritability of milk yield at different herd means and variances for production. J. Dairy Sci. 68:2737.
7 Mirande, S. L., and L. D. Van Vleck. 1985. Trends in genetic and phenotypic variances for milk production. J. Dairy Sci. 68:2278.

APPENDIX

Consider the linear model

$$y = Xb + Zu + e$$

and

$$v(y) = ZAZ'\sigma_u^2 + I\sigma_e^2$$

where y, b, u are vectors of the observations, fixed effects (here HYS), and sire effects, respectively; X, Z are the known incidence matrices for the fixed and random effects; and A is the numerator relationship matrix.

Famula (3) showed, using Henderson's selection model (4), the expectation of the reduction in SSS after fitting HYS in his formula [11], when selection had been practiced on a vector of herd means. This expectation is (the notation has been changed slightly)

$$\begin{aligned} E_s[R(u|b)] = {} & \text{trace}[Z'MZA]\ \sigma_u^2 \\ & + \text{trace}[Z'MZ(Z'MZ)^{-}]\ \sigma_e^2 \\ & - \text{trace}[Q'ZAZ'MZAZ'QH_0] \\ & \sigma_u^2 + (t'Q'ZAZ'MZAZ'Qt)\ \sigma_u^2 \end{aligned} \qquad [A1]$$

with

$$\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}',$$
$$\mathbf{Q} = (\mathbf{P}'\mathbf{X}'\mathbf{X}\mathbf{P})^{-1}\mathbf{P}'\mathbf{X}', \text{ and}$$
$$\mathbf{P} = \text{a matrix to link HYS to herds.}$$

Matrix $\mathbf{H}_0$ and vector t depend on the selection process but are not needed explicitly for the proof.

The first two terms of Equation [A1] are the standard terms for the unconditional (= no selection) case. The last two terms may result in a bias in the estimated sire variance, because they depend on the unknown $\mathbf{H}_0$ and t. To prove that these terms vanish for balanced designs, it is sufficient to show that the matrix $(\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{M}\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X})$, which appears in both terms, reduces to a zero matrix.

There are h HYS; each HYS has m, and each sire within a HYS has p observations. The vector y is ordered according to sire within HYS; $\mathbf{J}_i$ is a square matrix of ones of order i, and $\mathbf{D}_j$ is a block diagonal matrix with each block a $\mathbf{J}_j$ submatrix. Let the sires be unrelated ($\mathbf{A} = \mathbf{I}$). Then

$$\begin{aligned}
&\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{M}\mathbf{Z}\mathbf{A}\mathbf{Z}'\ \mathbf{X} \\
&\quad = \mathbf{X}'\mathbf{Z}\mathbf{Z}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Z}\mathbf{Z}'\mathbf{X} \\
&\quad = \mathbf{X}'\mathbf{Z}\mathbf{Z}'\mathbf{Z}\mathbf{Z}'\mathbf{X} - \mathbf{X}'\mathbf{Z}\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{Z}'\mathbf{X} \\
&\quad = \mathbf{X}'\mathbf{D}_p\mathbf{D}_p\mathbf{X} - (1/m)\mathbf{X}'\mathbf{D}_p\mathbf{D}_n\mathbf{D}_p\mathbf{X} \\
&\quad = p\mathbf{X}'\mathbf{D}_p\mathbf{X} - (p^2/n)\mathbf{X}'\mathbf{D}_n\mathbf{X} \\
&\quad = p(npJ_h) - (p^2/n)(n^2J_h) \\
&\quad = \mathbf{0}.
\end{aligned}$$