RUNNING HEAD: REFERENCE SIRE SCHEMES

GENETIC GROUPS AND THE FEASIBILITY OF REFERENCE SIRE SCHEMES[1]

S.P. Smith, R.D. Scarth and B. Tier
Animal Genetics and Breeding Unit
University of New England
Armidale NSW 2351 AUSTRALIA

SUMMARY

A method of measuring linkage in sire reference schemes is described. The method consists of determining estimability and prediction error variance for linear contrasts when the assumed animal model includes genetic groups. Base animals in a herd are assigned a common genetic group and different herds correspond to different genetic groups. The set of linear contrasts used to determine linkage is the set of group differences. Groups differences can be scrutinized using a simplified lease-squares analysis.

Key Words: animal model, genetic groups, linkage, sire reference schemes.

I. INTRODUCTION

Reference sire schemes are becoming increasingly popular for use in across-herd/flock genetic evaluations in livestock populations not making extensive use of artificial insemination. Practical animal breeders recognize the importance of adequate linkage in these reference sire schemes. Unrelated sires cannot be compared statistically unless they have progeny in common contemporary groups or they have other indirect "links". Alternatively, purists[2] consider the concept of linkage ill-founded because the mixed model allows for estimation of random effects independently of linkage. Even with little linkage it is possible to accurately estimate the difference between two

[1] This paper was written in 1988. Only small editorial improvements have been made in 2018, and no attempt was made to introduce materials from after 1988. However, FOULLEY, HANOCQ and BOICHARD (1992) provide some followup information, and this citation was added to the references for readers that want more.

[2] The purist is the statistician that takes the assumed model as true, and remains unwilling to look at assumptions beneath the model.

random sire effects; by assumption the sire effects are deviated around zero. Purists consider prediction error variance as the only criterion for whether sires are adequately compared or inadequately compared.

Practical concerns about linkage imply the existence of genetic groups (FOULLEY et al., 1988). That is, addressing fears about linkage is like saying, “what if the model is wrong because it has not accommodated genetic groups”. With genetic groups in the model, the purist approach can be adopted so as to vindicate the appled geneticists; as this paper shows with the animal model.

II. MODEL

We consider the animal model:

$$\mathbf{y}=\mathbf{X}\mathbf{b}+\mathbf{Z}\mathbf{a}+\mathbf{e} \qquad (1)$$

where **y**, **b**, **a** and **e** are vectors of observations, fixed effects, animal effects and residuals; **X** and **Z** are incidence matrices that assign various effects to observations. To complete the model’s specifications, first and second moments for **a** and **e** are needed. The usual declarations are:

$$\mathrm{E}(\mathbf{a})=\mathbf{0} \text{ and } \mathrm{Var}(\mathbf{a})=\mathbf{G} \qquad (1a)$$

$$\mathrm{E}(\mathbf{e})=\mathbf{0} \text{ and } \mathrm{Var}(\mathbf{e})=\mathbf{R} \qquad (1b)$$

Note that model (1) with companion clauses (1a) and (1b) is general enough to represent both univariate and multivariate models: for the univariate case $\mathbf{G}=\sigma_a^2\mathbf{A}$ where σ_a^2 is the additive genetic variance and **A** is the numerator relationship matrix.

To estimate **b** and **a** we solve HENDERSON’s (1973) mixed model equations:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \qquad (2)$$

where $\hat{\mathbf{b}}$ is the best linear unbiased estimate (BLUE) of **b** and $\hat{\mathbf{a}}$ is the best linear unbiased prediction (BLUP) of **a**.

If assumption (1a) is correct, then information about linkage is redundant and prediction error variances can be constructed from the inverse of the coefficient matrix of (2).

Alternatively, if we must worry about linkage then we necessarily reject assumption (1a) and need to construct useful alternatives.

III. ALTERNATIVE MODELS

Our treatment depends on technical machinery due to QUAAS (1979, personal communication).[3] He used these tools to construct a trivial proof of HENDERSON's (1975) formulae for $\mathbf{A}^{-1}$; see Hudson (1986) for illustration. Recently, GRASER et al. (1987) used the QUAAS machinery to accommodate fixed animals. These tools provide a simple and yet rigorous confirmation of the work of ELZO (1986) and, as this paper shows, WESTELL et al. (1988).

With the QUAAS machinery assumption (1a) can be expressed by

$$\mathbf{a}=\mathbf{Pa}+\mathbf{s} \tag{3}$$

where: $\mathbf{P}$ is a matrix such that $\mathbf{Pa}$ represents a vector of average mid-parent values (if both parents are known), ½ times parental effects (if one parent is known), and zeros (if no parent known); and $\mathbf{s}$ is a vector of residuals, which are attributed to segregation and random mating. To complete the assumptions for the univariate case, we define the diagonal matrix $\mathbf{D}$=Var($\mathbf{s}$) using principles of inheritance and stipulate that E($\mathbf{s}$)=$\mathbf{0}$. For the multivariate case, $\mathbf{D}$ is block diagonal or some permutation thereof.

With the distributional properties of $\mathbf{s}$ unspecified, (3) is the correct description of the biology. We may superimpose on the biology additional conditions: treat come of the elements of $\mathbf{a}$ as fixed (GRASER et al., 1987); introduce genetic groups, i.e., E($\mathbf{s}$)=$\mathbf{Qg}$ (WESTELL et al., 1988); or allow for different genetic parameters for different subpopulations (ELZO et al., 1986). Whatever condition we impose, we must be consistent with (3) and the related biology.

We are mainly concerned with genetic groups and hence we are interested in models of the form

$$\mathbf{a}=\mathbf{Qg}+\mathbf{Pa}+\check{\mathbf{s}}, \tag{4}$$

where $\mathbf{g}$ is a vector of fixed or random group effects, $\mathbf{Q}$ is a matrix that assigns groups to animals and $\check{\mathbf{s}}$ substitutes for $\mathbf{s}$-$\mathbf{Qg}$. To be consistent with biology, groups can only be

[3] His machinery was actually being readied for the publication in QUAAS (1988), while our paper was being written independently.

assigned to animals with unknown parents: usually a ½ of one group when one parent is unknown; and a whole group when both parents are unknown. Animals not assigned a group via **Q** are implicitly given their parents's group(s) via **P**. Animals that are not associated with a group directly or through relationships represent a "catch-all" group. Because the catch-all group effect is not in the model it is implicitly constrained to zero.

Now we must choose an appropriate matrix **Q** so as to model our worst fears about linkage. There is no unique way of doing this and it is the responsibility of the practitioner to come up with a suitable **Q**. Most of the concern about linkage is focused on comparing bulls used in herd A with different bulls used in herd B. This implies that all base animals within herd are assigned a common group and this is our proposal if nothing else is known. Different herds then get different groups, but this maybe too fine a grouping strategy when data can only justify a coarser grouping. Nevertheless, a finer grouping strategy may be also necessary if within herd comparisons are suspect, e.g., if a rancher buys a group of heifers from an outside source. Refinements can be extended to the point where individual animals are treated as independent variables possessing subjective variation. Assigning base animals to groups remains non-trivial.

In the above recommendation it may be necessary to devise a separate grouping strategy for those immigrant sires that do not come from recognized herds. Bulls that are brought in from over-seas, for example, may be grouped on year of entry into a system and country of origin.

IV. MODIFIED MIXED MODEL EQUATIONS

From (4) we obtain expression for **a** which is a function of **g** and **š**:

$$\mathbf{a}=(\mathbf{I}-\mathbf{P})^{-1}\mathbf{Qg} + (\mathbf{I}-\mathbf{P})^{-1}\mathbf{\check{s}}$$

Now substituting **a** into model (1) we obtain THOMPSON's (1979) group model:

$$\mathbf{y}=\mathbf{Xb} + \mathbf{Z}(\mathbf{I}-\mathbf{P})^{-1}\mathbf{Qg} +\mathbf{Zu} + \mathbf{e} \qquad (5)$$

where $\mathbf{u}=(\mathbf{I}-\mathbf{P})^{-1}\mathbf{\check{s}}$. Note that with **g** fixed, $\mathrm{Var}(\mathbf{a})=\mathrm{Var}(\mathbf{u})=\mathbf{G}= (\mathbf{I}-\mathbf{P})^{-1}\mathbf{D}(\mathbf{I}-\mathbf{P}')^{-1}$.[4]

Treating **g** fixed, the mixed model equations for (5) are

[4] The grouping model leaves the segregation variances unchanged, i.e., Var(**s**)=Var(**š**)=**D**.

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z(I-P)^{-1}Q} & \mathbf{X'R^{-1}Z} \\ \mathbf{Q'(I-P')^{-1}Z'R^{-1}X} & \mathbf{Q'(I-P')^{-1}Z'R^{-1}Z(I-P)^{-1}Q} & \mathbf{Q'(I-P')^{-1}Z'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z(I-P)^{-1}Q} & \mathbf{Z'R^{-1}Z+G^{-1}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{g}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{Q'(I-P')^{-1}Z'R^{-1}y} \\ \mathbf{Z'R^{-1}y} \end{bmatrix} \quad (6)$$

Now define the QUAAS and POLLAK (1981) transformation matrix

$$\mathbf{T} = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & -\mathbf{Q'(I-P')^{-1}} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix} \quad .$$

Multiplying both side of (6) by **T** and inserting $\mathbf{I}=\mathbf{T}'(\mathbf{T}')^{-1}$ between the coefficient matrix and the solution vector, we get the equations[5] of WESTELL et al. (1988):

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{0} & \mathbf{X'R^{-1}Z} \\ \mathbf{0} & \mathbf{Q'D^{-1}Q} & -\mathbf{Q'D^{-1}(I-P)} \\ \mathbf{Z'R^{-1}X} & -\mathbf{(I-P')D^{-1}Q} & \mathbf{Z'R^{-1}Z+G^{-1}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{g}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{0} \\ \mathbf{Z'R^{-1}y} \end{bmatrix} \quad (7)$$

The quantities $\mathbf{Q'D^{-1}Q}$, $-\mathbf{Q'D^{-1}(I-P)}$ and $\mathbf{G^{-1}}=\mathbf{(I-P')D^{-1}(I-P)}$ can be evaluated easily and simple rules can be devised for their construction (WESTELL et al., 1988). HENDERSON (1985) provides a different treatment of genetic groups.

If genetic groups are assumed random, (7) can be modified by augmenting the group by group part of the coefficient matrix with $\mathrm{Var}(\mathbf{g})^{-1}$. Moreover, some groups can be treated as fixed and others random.

Tp predict animal effects in the presence of group influences we simply solve (7). However, to determine whether linkage is adequate is more difficult. We agree with FOULLEY et al. (1988) in that it is more reasonable to check if herds in this case, i.e., groups more generally, are adequately compared rather than individual animals. Group 1 and group 2 are adequately compared if the prediction error variance of $\hat{g}_1$-$\hat{g}_2$ is sufficiently small. To find the prediction error variance requires certain inverse elements

[5] QUAAS (1988) formulated these same equations using an elegant alternative derivation.

of the coefficient matrix of (7). Presumably, we may employ sparse matrix absorption (TIER and SMITH 1989) to eliminate all non-group equations; the resulting matrix can then be inverted directly. However, this step is still liable to be difficult. A second concern is that some group contrasts may not be estimable, and we need to identify those contrasts. Estimable contrast can be established by following SEARLE (1971, p 185).[6]

V. RECOMMENDATION

Rather than working with (7) directly we propose a preliminary analysis to: determine linkage, and estimate the size of group effects. This is accomplished by a series of univariate least-squares analyses of model (5) ignoring **u**. Given that **b** is a vector of contemporary group effects, the univariate model is:

$$y_{ij}=b_i+\mathbf{w}_{ij}'\mathbf{g}+e_{ij} \qquad (8)$$

where y_{ij} is the j-th observation in the i-th contemporary group; $\mathbf{w}_{ij}$ is a vector of weights that equal the proportional contributions of each genetic group to y_{ij}. The $\mathbf{w}_{ij}$ can be computed by recursion in much the same way that **A** is evaluated by the tabular method; if xy and rs identify the parents of ij then $\mathbf{w}_{ij}=½\mathbf{w}_{xy}+½\mathbf{w}_{rs}$.

The normal equations for genetic groups with contemporary groups absorbed can be formed with one pass through the data. We determine estimability of group contrasts and scrutinize precision. This strategy reflects the belief that linkage occurs when animals associated with different genetic groups are compared in a common contemporary group.

Given that linkage is determined adequate, we next consider the size of the estimable group contrasts. If these contrasts are considered too large too ignore then we should return to system (7) for the main analysis.

After the main analysis, a criterion is needed for publishing individual breeding values. Our criterion is to publish a set of breeding values if: the associated groups are well estimated and linked; the breeding values have a sufficient accuracy.

VI. CONCLUSION

Whereas our recommendations is computationally easy, it can be criticized because a least-squares analysis of (8) is not fully efficient. However, it is difficult to see how

[6] A practical way to determine estimablity of group contrasts in a large system is to seek two or more iterative solutions, but using different starting values for the group effects. The contrast are estimable if they remain invariant to the starting values.

decisions resulting from (8) would be far removed from decisions resulting from (5). For example, a linear contrast of fixed effects is estimable under (5) if and only if it is estimable under (8) (FOULLEY et al., 1988). If the practitioner feels that a more exact analysis is justified then a closer examination of (5) and (7) is warranted; there are other ways to approximate prediction error variances than our least-squares approach.

One mistaken view is that treating part of the genetic component as fixed is inherently bad; to be conservative we should regress effects back to zero. However, we cannot at one hand complain about linkage and then argue that there are no fixed effects. Linkage is a problem, precisely because our true model possesses fixed effects. If it is not possible to estimate certain fixed effects accurately, then we should not publish the associated breeding values. If animal breeder must regress breeding values back to zero, then they should at least acknowledge the existence of genetic groups and treat these groups as random. At least we can haggle with the subjectivity involved in determining group covariance structure. The undeclared subjectivity associated with the incorrect use of (2) is untouchable.

It may be necessary to modify our procedure, i.e., system (7), when selected sires with prior information are brought into the evaluation program. This can be accomplished by the approach of GODDARD and SMITH (1988). This method involves: setting up the prediction equations treating selected sires as fixed; then using the prior information in the same way that HENDERSON (1984, p 102) uses external information on fixed effects.

Lastly, close analysis may reveal that traditional reference sire schemes are not as useful in establishing links as we would like them to be. That is, our purists instincts may be way too optimistic. If this is true then modification may be required to increase the use of reference sires to improve linkage.

REFERENCES

ELZO, M.A., 1986, Inverse of single trait additive genetic covariance matrix with unequal variances across additive genetic groups. J. Dairy Sci., 59, 569-74.

FOULLEY, J.L., J. BOUIX, B. GOFFINET and J.M. ELSEN, 1988, Connectedness in genetic evaluation. In: International Symposium, Advances in Statistical Methods for Genetic Improvement of Livestock, Editors D. Gianola and K Hammond, Heidelberg, Springer-Verlag.

FOULLEY, J.L., E. HANOCQ and D. BOICHARD, 1992, A criterion for measuring the degree of connectedness in linear models of genetic evaluation. Gene. Sel. Evol., 24 , 315-30.

GODDARD, M., and S.P SMITH, 1988, Utilizing overseas information in the estimation of breeding values. Unpublished memo.[7]

GRASER, H.-U., S.P. SMITH and B. TIER, 1987, A derivative-free approach for estimating variance components in animal models by restricted maximum likelihood. J. Animal Sci., 64, 1360-70.

HENDERSON, C.R., 1973, Sire evaluation and genetic trends. In: Proceedings of the Animal Breeding and Genetics Symposium in Honor of Dr J.L. Lush, pp.10-41, American Society of Animal Science: Champaign, Illinois.

HENDERSON, C.R., 1975, Rapid method for computing the inverse of a relationship matrix. J. Dairy Sci., 58, 1727-30.

HENDERSON, C.R., 1984, Application of Linear Models in Animal Breeding. 462 p., University of Guelph.

HENDERSON, C.R., 1985, Best linear unbiased prediction using relationship matrices derived from selected base populations. J. Dairy Sci., 68, 443-8.

HUDSON, G.F.S., 1986, Computing genetic evaluations through application of generalized least squares to an animal model. Genet. Sel. Evol., 18, 13-40.

QUAAS, R.L., and E.J. POLLAK, 1981, Modified equations through for sire models with groups, J. Dairy Sci., 64, 1868-72.

QUAAS, R.L., 1988, Additive genetic model with groups and relationships. J. Dairy Sci., 17, 1334-1345.

SEARLE, S.R., 1971, Linear Models. 532 p., New York, John Wiley and Sons, Inc.

THOMPSON, R., 1979, Sire evaluation. Biometrics, 35, 339-53.

TIER, B., and S.P. SMITH, 1989, Use of sparse matrix absorption in animal breeding. Genet. Sel. Evol., 21,457-66.

WESTELL, R.A., R.L. QUAAS and L.D. VAN VLECK,1988, Genetic groups in an animal model. J. Dairy Sci., 71(5), 1310-18.

[7] Goddard's broad views can be found in "International Genetic Evaluation," in the Proceeding of the Australian Association of Animal Breeding and Genetics, 10, 267-272.