SAMPLING OF TRAITS AND DATA IN ESTIMATION OF GENETIC PARAMETERS

Just Jensen
National Institute of Animal Science,
Foulum, 8830 Tjele, Denmark

and

Terri L. Moore and Ivan L. Mao
Department of Animal Science
Michigan State University
East Lansing, 48824, U.S.A.

SUMMARY

Results from simulation studies on REML estimation of genetic parameters from sampled traits in unselected populations and data in population undergoing selction are presented. Inclusion of correlated traits in a multiple trait analysis in unselected populations did not increase accuracy of heritability estimates, but did increase accuracy of estimates of genetic correlations. REML estimates of heritability were unbiased in unselected populations if all data and all relationships were incorporated in the model. If only recent data were included, estimates were still unbiased if all relationships were taken into account.

INTRODUCTION

Genetic parameters are often estimated from large datasets including information on several traits. Inclusion of all data and all traits simultaneously in a multiple trait analysis is often computationally difficult. Therefore data are often sampled by including only a subset of the traits of interest in each analysis, and by including only a subset of the animals with records. For example, the subset of animal records could be from a certain time period.

Field records used for parameter estimation usually come from commercial populations that undergo intense selection for one or more traits. The genetic (co)variances would change over time due to accumulation of inbreeding and gametic phase disequilibrium (Bulmer, 1971). In order to draw genetic inferences about the population, parameters prior to selection must be known.

Schaeffer (1987) showed that for certain translation invariant selection rules REML estimates are not biased by selection if all data used in selection decisions are included in the analysis. In many practical situations, however, the selection rules were not translation invariant and often only recent data are available for analysis, but with pedigree information available back to the base population.

The objective of this paper is to present results of different strategies of selecting subsets of traits in unselected populations or subsets of data in time periods in populations undergoing selection.

MATERIAL AND METHODS

Subset of Traits

The effect of number of traits included in an analysis was investigated for data from an unselected population, i.e. sampling according to traits. Data were simulated for four traits that were normally distributed with varying heritabilities and genetic correlations. Within a situation all genetic correlations were identical and residual correlations were kept constant at .5 for all situations. A total of 12 situations with varying (co)variance structure, as indicated in Table 1 and 2, were investigated. Each situation was replicated 50 times and in each replicate, data for 40 sires with an average progeny group size of 20 were simulated. Progeny were assigned to 22 management groups such that each sire had progeny in four management groups. The data in each replicate were analyzed with a multiple trait EM-REML (Expectation Maximization-Restricted Maximum Likelihood) procedure using canonical and Householder transformations as described by Jensen and Mao (1988). All possible four, three, two and single trait analyses were performed on each dataset such that a total of 56 parameters were estimated per dataset.

Subset of Data Over Time

Various strategies for sampling data according to time in populations undergoing selection for one or more traits were investigated. A stochastic model was used to simulate data for dual purpose populations selecting for milk and beef, and to simulate dairy populations selecting for milk only. The population size for a dual purpose population was 200 cows per year and there was 1000 cows per year for a dairy population. Beef traits were assumed to be measured on a central testing station for future AI bulls and milk production was assumed to be measured in commercial herds. The two selection schemes were simulated over a 15 yr time horizon, and were replicated 15 times.

Each run of the simulation model generated a data file and a pedigree file. Data were sampled from each set of data according to three schemes. In Scheme 1 all data, and the full pedigree file were used in the analysis. In Scheme 2, only data from the last 5 yr were included, but the full pedigree file was used for tracing relationships. Scheme 3 also used records from the last 5 yr only and pedigree information only from the last 5 yr.

Each sample of data was analyzed by two multitrait models. A full two-trait animal model, and a model where the submodel for growth was an animal model but the submodel for milk was a sire model. Such a model was possible since sires had records on beef themselves and their female progeny had records on milk production. Genetic parameters were estimated by a derivative free REML algorithm similar to the one described by Meyer (1990).

RESULTS

Subset of Traits

Average bias and root mean square error (RMSE) were computed for all genetic parameter estimates. The number of biased estimates found 'was no different from the number expected based on the level of significance used.

The RMSE of heritability estimates is shown in Table 1 for a few selected situations. The RSME of estimates of genetic parameter estimates is very dependent on the underlying true parameters. For heritability estimates, RMSE were not reduced by increasing the number of traits in the analysis, i.e. there were no advantage of including correlated traits when estimating heritabilities in the situations investigated. The RMSEs of estimates of genetic correlations are shown in Table 2 for a few selected situations. Now, there seemed to be an advantage of including more traits in the analysis, since the RMSE decreases with increasing number of traits included. The reason seems to be that the (co)variance matrices estimated must be positive definite in order to be EM-REML estimates. Thus, by increasing the number of traits in the analysis, more severe restrictions are imposed on the sample (co)variance matrices. This could also explain why the effect of including more traits is greater when the (co)variance matrix to be estimated is close to the edge of the parameter space. For example, the genetic correlation, $r_A = -.2$ vs. $r_A = .2$, since r_A cannot be below -.31 for four equally correlated traits. A stronger negative correlation would indicate a non-positive definite (co)variance matrix.

Table 1. Average root mean square error (RMSE) of heritability (h^2) estimates in selected different levels of h^2 and genetic correlation (r_A) and number of traits in analysis.

No. of traits in analysis		h^2=.2				h^2=.6			
	r_A:	.2	.8	-.2	-.3	.2	.8	-.2	-.3
4		.09	.09	.09	.08	.15	.13	.16	.15
3		.09	.10	.09	.08	.15	.13	.16	.15
2		.09	.10	.09	.09	.15	.13	.16	.15
1		.09	.10	.10	.08	.15	.13	.16	.15

Table 2. Average root mean square error(RMSE) of genetic correlation estimates in selected different levels of h^2 and genetic correlation (r_A) and number of traits in analysis.

		r_A=.2			r_A=-.2		
No. of traits in analysis:		4	3	2	4	3	2
h^2							
.2	.2	.35	.38	.38	.32	.35	.39
.1	.8	.35	.38	.41	.35	.37	.40
.3	.8	.23	.23	.23	.26	.27	.27
.6	.8	.17	.17	.17	.18	.19	.19
.8	.8	.20	.20	.20	.20	.20	.20

Subset of Data Over Time

Results are shown in Table 3. Use of the full animal model with all data and all relationships yielded unbiased estimates of heritability of milk production. This was also the case if only late data were included, but all relationships were traced back to the base population. Most of the estimates of heritability of growth were biased upwards, which was an unexpected result and we have no explanation. Use of a sire submodel for milk production gave unbiased results in the dual purpose populations if all data and all relationships among males were incorporated. Inclusion of records for growth on the sires themselves seemed to alleviate bias due to selection. The use of a sire model in the single purpose dairy populations in all cases yielded biased heritability estimates.

Table 3. Estimates of heritability, averaged over 15 replicates with SE in parenthesis, in cattle populations undergoing selection.

	Sampling	Dual purpose		Dairy
Model	scheme	milk	growth	milk
(True parameters		.25	.50	.25)
Full Animal model	1	.25 (.06)	.64(.17)	
	2	.25 (.09)	.62(.22)	
	3	.20[a](.10)	.50(.24)	
Sire model for milk	1	.25 (.12)	.78(.12)	.20[a](.08)
Animal model for growth	2	.21 (.17)	.48(.26)	.20[a](.08)
	3	.21 (.18)	.60(.21)	.19[a](.08)

[a] Significantly different from true parameters at 5% level.

CONCLUSION

Inclusion of correlated traits in multiple trait analysis did not increase precision of estimates of heritability in unselected populations, but increased precision of estimates of genetic correlations.

Use of a full animal model with complete relationships yielded unbiased estimates of genetic parameters in populations undergoing selection. Use of sire models for milk production seems to yield unbiased results if records on growth of the bulls themselves were included in the model as long as all data were included in the analysis. Use of sire models in dairy populations always yielded biased estimates.

REFERENCES

BULMER, M.G. 1971. Amer. Nat. 105:201-211.
JENSEN; J. and MAO, I.L. 1988. J. Anim. Sci. 66:2750-2761.
MEYER. K. 1990. Genet. Sel. Evol. (Submitted).
SCHAEFFER, L.R. 1987. J. Dairy Sci. 70:661-671.
MOORE, T.L., MAO, I.L. and JENSEN, J. 1989. J. Dairy Sci. 72 (Suppl. 1):32
JENSEN, J. and MAO, I.L. 1989. J. Dairy Sci. 72 (Suppl. 1):33.