CAN BLUP AND REML BE IMPROVED UPON?

Daniel Gianola
Department of Animal Sciences, University of Illinois,USA

SUMMARY

Nonlinear, biased estimators and predictors of the James-Stein form are compared numerically with ML estimation of fixed effects and BLUP prediction. At least in some situations considered, it appears that considerable improvement can be achieved in terms of conditional mean squared error and conditional absolute error. The possibility of improving upon REML using Bayesian methods is raised. Methods for deriving the exact posterior distribution of heritability in an animal model and either exact or approximate marginal posterior distributions of variance components are discussed.

INTRODUCTION

Any casual review of the animal breeding literature will reveal that best linear unbiased estimation and prediction (BLUE and BLUP) of fixed effects and random variables, respectively, and restricted maximum likelihood estimation of dispersion parameters (REML), dominate the methodological scene. Many reasons can be advanced for this. In some instances the assumption of normality required by REML (and implicitly by BLUE and BLUP) is manifestly violated and nonlinear methods are required; other papers in this volume document these. However, is it possible to go beyond BLUP and REML in the case of normally distributed traits? We believe the answer is yes and the objective of this paper, is to provide some preliminary analytical and numerical results supporting this contention. For reasons of space, theory, details and references will be restricted to a minimum. Much more research is needed before the methods discussed here can be recommend unequivocally for general purposes.

The authors subscribe to the Bayesian viewpoint and feel that a large part of future methodological developments will either be inspired by this "school", or will have a Bayesian justification. However, the position taken in this paper is eclectic, as some results applying to performance over repeated sampling are reported.

THE JAMES-STEIN PHENOMENON IN ANIMAL BREEDING

We consider the balanced model:

$$y_{ijk}=g_i+s_{ij}+e_{ijk} \qquad (1)$$

where the g's are fixed group effects (i=1,...,G), $s_{ij}\sim NIID(0,\sigma_s^2)$ with j=1,..,S for every i, $e_{ijk}\sim NIID(0,\sigma_e^2)$ with k=1,...,n for every (i,j). The sire effects s_{ij} and the residuals are mutually independent. The following estimators and predictors of g and s, respectively, are studied:

1) $ML(g_i)= E_i= \sum_{j=1}^{S} \bar{y}_{ij} /S$ $\quad BLUP(s_{ij})= P_i= n(n+\alpha)^{-1}(\bar{y}_{ij}-E_i)$ $\quad T_i=E_i+P_i$

Above, $\bar{y}_{1j}$ denotes daughter average and $\alpha = \sigma_e^2/\sigma_s^2$.

2) $E_2 = [1-(G-2)v_\epsilon/(Sxk)]E_1$ $\quad P_2 = n(n+\alpha)^{-1}(\bar{y}_{1j}-E_2)$ $\quad T_2=E_2+P_2$

Here $v_\epsilon = \sigma_s^2+\sigma_e^2/n$, $k=\sum_{i=1}^{G} E_i^2$ and E_2 is the James-Stein (1961) estimator of g_i.

3) $E_3 = [1- Gv_\epsilon/(S \Sigma g_i^2+Gv_\epsilon)]E_1$ $\quad P_3=n(n+\alpha)^{-1}(\bar{y}_{1j}-E_3)$ $\quad T_3=E_3+P_3$

Estimator 3 has minimum mean squared error in the class of linear estimators of group effects (Theil, 1971), but it requires knowing these! It was used as a sort of "control".

4) As above, but with g_i replaced by its maximum likelihood estimator E_1. We refer to E_4, P_4 and T_4 as "estimated Theil" estimators and predictors.

5) $E_5 = \bar{g}+\{1-Gv_\epsilon/[S\sum_{i=1}^{G}(E_1-\bar{g})^2+Gv_\epsilon]\}(E_1-\bar{g})$ $\quad P_5=n(n+\alpha)^{-1}(\bar{y}_{1j}-E_5)$ $\quad T_5=E_5+P_5$

The estimator E_5 is in the form of the minimum mean squared error estimator E_3 but shrinkage is towards the average group effect, rather than towards zero. The statistics are denoted as "modified Theil".

6) The estimators and predictors have a form similar to (5) above, but with the average group effect replaced by its maximum likelihood estimator. We refer to these as "Lindley" statistics.

The logic for the choice of estimators is based on James and Stein (1961) who found that that the maximum likelihood estimator of the location vector in a linear model (normality assumed) is inadmissible under squared error loss. They proposed the orthonormal version of estimator E_2, which was shown to dominate the maximum likelihood estimator (for $G>2$). Although the James-Stein estimator has been advocated in terms of its sampling properties, it has a Bayesian interpretation and inspiration (Zellner and Vandaele, 1974). Because it can be shown that BLUP(s_{ij}) cannot be improved upon in a mean squared error sense (this holds unconditionally , e.g., over repeated samples of sires, but not conditionally on the sires being evaluated), we also considered predictors T_i (i=2,...,6) with E_1 in lieu of E_i in the "sire" portion of the predictor.

The estimators and predictors under consideration were studied in the following way. Sire and residual effects were generated from independent normal distributions with null means, and variances as outlined below. Group effects were fixed, and the difference between last and first group effect was set to be equal to two genetic standard deviations. Only one sample was drawn from the distribution of sires, whereas the number of drawings from the residual distribution depended on the number of progeny per sire. In this respect, all estimators and predictors fall in the class of empirical Bayes (Efron and Morris, 1973) where the items are of interest by themselves, as opposed to the interest being centered on the population from which samples are drawn. Empirical Bayes exploits the "commonality" among items, but the sampling properties are evaluated conditionally upon the sample (of sire effects in this

case) drawn. With the true group and sire effects stored, we computed conditional average mean squared error and absolute error for each of the estimators studied.

The situations reported here are as follows: A) "Low heritability-low variance" G=20, S=4, n=5, heritability=.01 (α=399), residual variance=10, base group effect=50. This could represent the case of a fitness trait. B) "High coefficient of variation", such as litter size in pigs. Here G=20, S=3, n=5, heritability= .10 (α=39), residual variance= 5, base group effect= 8. C) "Moderate heritability" (e.g., milk yield in dairy cattle), with G=15, S=2, n=20, heritability=$\frac{1}{4}$(α=15), residual variance= 250000, base group effect= 4500. D) In order to examine the effect of using E_1 instead of the appropriate E_i in T_i (i=2,..., 6), the following parameters were employed: G=5, S=2, n=10, α=15, residual variance= 1000, and base group effect = 1000.

Conditional mean squared and absolute errors of prediction are in Tables 1 and 2. The values are relative to those of ML and BLUP. It should be kept in mind that the Theil and estimated Theil (E-Theil) statistics are not feasible in practice, because knowledge of the fixed effects is required for their calculation. For g+s, results presented are based on using BLUP for the "sire" part of the predictor and the corresponding estimator for the "group" portion.

Table 1 Relative conditional mean squared errors of alternative estimators of fixed group effects and of predictors of breeding value

	"Low heritability"			"High coefficient of variation"			"Moderate heritability"		
Statistic	g	s	g+s	g	s	g+s	g	s	g+s
ML+Blup	100	100	100	100	100	100	100	100	100
Stein	101	100	101	94	100	95	99	99	101
Theil	101	100	101	93	100	94	98	99	101
E-Theil	101	100	101	93	100	94	98	99	101
M-Theil	28	100	33	37	97	49	66	89	107
Lindley	12	99	20	41	98	54	70	91	112

Table 2 Relative conditional absolute errors of alternative estimators of fixed group effects and of predictors of breeding value

	"Low heritability"			"High coefficient of variation"			"Moderate heritability"		
Statistic	g	s	g+s	g	s	g+s	g	s	g+s
ML+Blup	100	100	100	100	100	100	100	100	100
Stein	100	100	100	96	100	98	100	100	100
Theil	100	100	100	96	100	97	100	100	100
E-Theil	100	100	100	96	100	97	100	100	100
M-Theil	54	100	57	62	99	72	82	93	103
Lindley	35	100	46	69	100	77	81	94	106

At least within the range of the situations considered, it is possible to obtain a sizable improvement over ML for estimation of group effects, and over BLUP for prediction of group+sire values, but not for sire effects. For example, at "low heritability", the mean squared error of the Lindley estimator of group effects is less than an eighth of that of ML; the absolute error of the Lindley predictor of g+s effects is less than half of the one for BLUP. For the "moderate heritability"

situation, the improvement is still sizable for estimation of group effects, but nil or negative for prediction of breeding value. Concerning results for situation D, using the corresponding estimator in the "sire" portion of the g+s predictor increased mean squared (as expected) and absolute error for all statistics evaluated.

In summary, it appears that BLUE (ML under normality) and BLUP can be improved upon both in a mean squared error and absolute error sense. The most promising statistic is the one termed "Lindley" here; it consists of shrinking estimated fixed effects to their average estimated value. There are many Bayesian and non-Bayesian ways that this can be accomplished, which suggests room for further study. As it always happens with the sampling school of statistics, the results are highly dependent on the part of the parameter space that one is situated (Peixoto and Harville, 1986) and no generalizations can be made. Although we presented results based on drawing a single sample of sire effects, we conjecture that conclusions based on, say, unconditional mean squared errors, should be the same. This is so because the conditional end-points evaluated here were obtained after averaging over all sires in the sample. After all, the unconditional end-points are the expected values of the conditional ones, the expectation taken over the distribution of sire effects. In conclusion, ML estimation of fixed effects and genetic evaluation by BLUP can be improved upon, at least in some instances. Whether "closer" estimates and predictions lead to faster genetic progress is an open question.

BEYOND REML

From a Bayesian perspective, REML estimation of variance components is "better" than ML because the fixed effects (appearing as nuisance parameters) are integrated out. What happens if some of the variances are nuisances parameters themselves? From a likelihood inference viewpoint it seems impossible to factorize a likelihood such that there is a portion depending only on the variances of interest. However, it is feasible to carry out such marginalization by either exact (numerical) or approximate (analytical) Bayesian analysis.

In a Bayesian context (consider flat priors, but this can be relaxed), it is possible to obtain the exact posterior density of the ratio of variance components (or of any function thereof), and the conditional density of any variance component, given the ratio. For example, in an animal model with two variance components, the posterior density of heritability (h) is:

$$p(h|\mathbf{y}) \propto (1-h)^{\frac{1}{2}(q-4)} h^{-\frac{1}{2}q} |\mathbf{C}|^{-\frac{1}{2}} [\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\theta}}'\mathbf{r}]^{-\frac{1}{2}(n-p-4)} \quad ; \; 0<h<1 \qquad (2)$$

where $\mathbf{y}$ is data, $\hat{\boldsymbol{\theta}}$ is the solution to the mixed model equations, $\mathbf{C}$ is the coefficient matrix, $\mathbf{r}$ is the vector of right hand-sides, p=rank ($\mathbf{X}$) and q is number of animals. Using data of Harville and Callanan (1990), we constructed the whole posterior distribution and, in particular, a highest posterior density region of 84% coverage was found to be .2-1.0. In contrast, the above authors used REML to develop an 80% confidence interval which covered the entire parameter space and even more! Thus, we can attain a sharper state of knowledge via Bayesian analysis.

With respect to inference about individual variance components, it can be shown that the conditional density of a variance given the ratios is in the inverted chi-square form. Hence, marginal means and variances can be obtained using the

moments of the conditional distribution (obtained explictly) and then integrating numerically with respect to the distribution of the ratios. An approximation that can be justified in many instances would be finding the mode of the density of the ratios, and then using the inverted chi-square densities evaluated at the modal values to describe uncertainty about individual variances. From, these, an approximate marginal posterior distribution can be generated. Whereas the distributions of sampling estimators of variance components are unknown even in the most trivial models, one can go much further in Bayesian analysis. Practicioners that accept the logic of REML ("taking into account degrees of freedom lost") should accept the idea of marginalizing likelihoods even further, as additional degrees of freedom stemming from the "nuisance" variance are also taken into account. For details, the readers are referred to Gianola and Foulley (1990), and Gianola et al. (1990).

As a final remark, it should be clear that there is room for further theoretical developments in statistical genetics. Although some of the new ideas involve more intensive computation, we are confident that some of the problems posed can be solved (even with currently existing hardware) using efficient numerical strategies.

REFERENCES

EFRON, B. and MORRIS, C. 1973. J. Amer. Stat. Assoc. 68: 117-130.

GIANOLA, D. and FOULLEY, J.L. 1990. Genetics, Selection, Evolution (submitted)

GIANOLA, D., FOULLEY, J.L. and FERNANDO, R.L. 1990. Theor. Appl. Genet. (submitted)

HARVILLE, D.A. and CALLANAN, T.P. 1990. In Gianola, D. and Hammond, K. (eds.): Advances in statistical methods for genetic improvement of livestock. Springer-Verlag, Heidelberg.

JAMES, W. and STEIN, C. 1961. Proc. Fourth Berkeley Symp. Math. Stat. and Prob. 1: 361-379

PEIXOTO, J.L. and HARVILLE, D.A. 1986. J. Amer. Stat. Assoc. 81: 431-436.

THEIL, H. 1971. Principles of Econometrics. Wiley, New York

ZELLNER, A. and VANDAELE, W. 1974. In Fienberg, S.E. and Zellner, A. (eds.): Studies in Bayesian Econometrics. North Holland, Amsterdam.