A COMPARISON OF BENDING AND RIDGE REGRESSION IN SELECTION INDEX ESTIMATION

A. M. SAXTON, USA
Department of Experimental Statistics, Louisiana State University
Baton Rouge, LA 70803

Summary

Two methods for genetic selection index estimation based on ridge regression are proposed and compared to 'bending'. Both ridge and bending methods attempt to increase efficiency by altering eigenvalues known to be biased when estimated from small experiments. While the ridge index showed improvement over the usual index, it was not better than bending, particularly for smaller experiments. One advantage of the ridge index was that best performance occurred for ridge parameters of approximately 2, whereas the optimum bending parameter varies widely depending mainly on experiment size and heritabilities. In practice, it may be hard to obtain the optimum bending parameter with great accuracy.

1. Introduction

The selection index is one of the most widely used techniques for the genetic improvement of populations. If genetic change is linearly related to economic value, an index is also the most efficient technique. A problem which has been addressed by several authors (Harris, 1964; Hayes and Hill, 1981) is that estimation of selection index weights must be based on estimates of phenotypic and genetic variances and covariances, since population values are never known in practice. The selection index has maximum efficiency when population values are used. The effect incorrect values have on efficiency has been examined by Williams (1962a,b) and Sales and Hill (1967a,b).

Hayes and Hill (1981) suggested a method, named bending, in which bias is introduced into the estimation procedure to get index weights which are closer to true values. This is an example of the statistical notion that although unbiasedness is a desirable property, unbiased estimates are not necessarily best in the sense of being close to population values. These authors indicated that bending is similar in concept to ridge regression. An exact analogy between regression and the selection index cannot be made, because although an index is essentially a prediction (or regression) equation, the quantity being predicted is the unobservable aggregate genotype. The purpose of this paper is to see if a ridge-like estimator can be developed which will give estimates closer to true values than bending. A substantial body of research exists for ridge regression (Hocking, 1976), so it would be convenient to be able to apply these results to selection index estimation.

2. Methods

Following the notation of Hayes and Hill (1981), an index $I = \underline{b}'\underline{x}$ for individual selection will be considered, with $\underline{b}$, the vector of index weights, to be estimated and $\underline{x}$ being the vector of observed phenotypic values. The $\underline{b}$ are chosen to maximize the correlation between I and $H = \underline{a}'\underline{g}$, the aggregate genotype, where $\underline{a}$ is the vector of economic weights and $\underline{g}$ is the vector of breeding values. As is well known, the solution is

$$\hat{\underline{b}} = \hat{\underline{P}}^{-1} \hat{\underline{Ga}}$$

where $\hat{\underline{P}}$ and $\hat{\underline{G}}$ are the observed phenotypic and genetic variance-covariance matrices, respectively.

Predicted response, with i representing selection intensity, is

$$\hat{R} = i\ (\hat{\underline{b}}'\hat{\underline{P}}\hat{\underline{b}})^{.5}$$

while actual response is

$$R^a = i\ \hat{\underline{b}}'\underline{G}\underline{a}\ (\hat{\underline{b}}'\hat{\underline{P}}\hat{\underline{b}})^{-.5},$$

which depends on the unknown population parameters $\underline{P}$ and $\underline{G}$. As in Hayes and Hill (1981), for simplicity assume $\underline{P}$ and $\underline{G}$ are estimated from a balanced one-way multivariate analysis of variance as follows,

Source	df	MS	EMS
Sires	s-1	$\underline{B}$	$\underline{P}$ - .25 $\underline{G}$ + .25 n $\underline{G}$
Progeny (Sire)	s(n-1)	$\underline{W}$	$\underline{P}$ - .25 $\underline{G}$

which yields estimates of $\underline{P}$ and $\underline{G}$ by setting MS equal to EMS.

Ridge regression estimation (Hoerl and Kennard, 1970) modifies the usual least squares equation

$$\underline{b} = (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{y}$$

by adding a constant to the diagonal of the $\underline{X}'\underline{X}$ matrix

$$\underline{b}_R = (\underline{X}'\underline{X} + k\ \underline{I})^{-1}\ \underline{X}'\underline{y}.$$

For the appropriate choice of k, this can result in $\underline{b}_R$ being closer to $\underline{b}$, the true values. There is some controversy about the general usefulness of ridge regression and also about how to choose k (Hocking, 1976).

If the similarity between $\underline{P}$ and $\underline{X}'\underline{X}$ is noted, one ridge-like estimator which might be proposed is

$$\hat{\underline{b}} = [\hat{\underline{P}} + k*\text{diag}\ (\hat{\underline{P}})]^{-1}\ \hat{\underline{G}}\underline{a}.$$

This estimator did not show any improvement over the usual index, so it will not be considered further. Another estimator which can be proposed modifies both $\underline{P}$ and $\underline{G}$, since both are estimated quantities

$$\hat{\underline{b}}_k = [\hat{\underline{P}}^{-1}\hat{\underline{G}} + k*\text{diag}\ (\hat{\underline{P}}^{-1}\hat{\underline{G}})]\ \underline{a}.$$

In these equations diag is the usual operator which zeros out all off diagonal elements. This is used so that changes on the diagonal elements are proportional to the original values.

The procedure of Hayes and Hill (1981) was used to test the performance of these estimators. Briefly, population values for $\underline{P}$ and $\underline{G}$ were chosen which had been transformed (Hayes and Hill, 1980) such that $\underline{P}$ was an identity matrix and $\underline{G}$ was a diagonal matrix of transformed heritabilities. Sample mean square matrices were generated from population values (Hill and

Thompson, 1978) and used to calculate $\underline{b}$. For sire numbers greater than 100, 500 samples were generated, but 1500 samples were used otherwise because of the greater variability in the results. To limit the number of simulations to a reasonable quantity, 8 combinations of 5 factors were selected using a fractional factorial design. Factors considered were number of traits, number of sires (s), number of progeny (n) and range in economic weights. Properties used to compare the various methods for calculating a selection index were

$E(R^a)/R$ = average fraction of possible genetic response actually achieved

and % improved = percentage of experiments in which the modified index gave at least as much response as the usual index.

3. Results

The performance of the suggested ridge index at various values of k is shown in Table 1, along with a comparison to optimum bending results. The ridge index gave higher expected gains than the usual index. In comparison to bending, ridge index did worse for s=30, but did almost as well for s=150. The percent improved values were often higher, also reflecting the good performance of the ridge index for s=150. The number of sires was by far the most important influence on the results, followed by size of heritabilities.

Ridge index is able to improve genetic gains for the same reason as bending. Both reduce the spread in the eigenvalues of the $\underline{P}^{-1}\underline{G}$ matrix, as shown in Table 2. For better illustration, extreme sire numbers are used. For s=10, the outward bias in the eigenvalues is expected to be large, so a bending parameter of about .8 is needed to remove the bias. The ridge method could not alter the eigenvalues to the extent required. For s=500 the bias is much smaller, and both bending and ridge parameters could be found which gave values close to the true values. The ridge method changed the eigenvalues relatively little, which may be advantageous for large sire numbers.

4. Discussion

Clearly bending is the method of choice. Although the ridge index could match bending for large sire numbers, in this situation modification of the index is not necessary. For small sire numbers bending significantly out performed the ridge method. This result was caused by the inability of the ridge method to change the eigenvalues enough to give index weights close to the true values. Bending, with its direct manipulation, gives total control over the eigenvalues and thus allows improved performance in all circumstances if the correct bending parameter is used.

The ridge index appears to correct one difficulty of the bending method, which is selection of the bending parameter. The optimum bending value depends on an unknown function of sire number, heritability, and other parameters and varies greatly as seen in Table 1 and in Hayes and Hill (1981). The ridge index always had best performance at k around 2, this being a compromise between increasing actual gain and decreasing % improve. More research is needed on how to best choose the bending parameter.

Table 1. Comparison of the ridge index to unmodified and optimum bending indexes.

h^2	a	s	n		usual index	bending	γ^a	ridge(k) .6	1.2	2.0	4.0
.1,.2	1,1	30	8	$E(R^a)/R$	.56	.80	.9	.63	.65	.66	.67
				% improve		65		56	55	55	54
.6,.8	1,10	30	8	$E(R^a)/R$	.97	1.0	1.	.98	.99	.99	1.0
				% improve		93		93	92	92	90
.1,.2, .15,.2,.3	1,1,1, 1,1	150	8	$E(R^a)/R$	.84	.96	.8	.91	.93	.93	.94
				% improve		89		92	90	88	86
.1,.15, .2,.2,.3	1,2,5, 8,10	30	20	$E(R^a)/R$	.83	.98	.9	.90	.92	.93	.94
				% improve		98		96	95	94	93
.6,.7,.8, .85,.9	1,1,1, 1,1	30	20	$E(R^a)/R$	.93	.99	.9	.96	.97	.97	.97
				% improve		98		92	91	89	88
.1,.2	10,1	150	20	$E(R^a)/R$	.960	.996	.9	.978	.985	.988	.991
				% improve		77		89	88	87	86
.6,.8	1,1	150	20	$E(R^a)/R$	.996	.997	.2	.997	.997	.997	.997
				% improve		51		52	52	51	51
.6,.7,.8, .85,.9	1,10,8, 2,5	150	8	$E(R^a)/R$	.974	.996	.8	.985	.988	.990	.991
				% improve		97		97	97	96	95

[a] optimum bending parameter

Table 2. Changes in the relative eigenvalues of the $\underline{P}^{-1}\underline{G}$ matrix for an individual sample. Heritabilities are .8 and .6 with both economic weights equal to 1 so the true relative eigenvalues are .8/.6 or 1/.75. Only the second eigenvalue is shown. In all cases the number of progeny is 500.

Bending Parameter	Bending 10 sires	Bending 500 sires	Ridge Parameter	Ridge 10 sires	Ridge 500 sires
0	.165	.731	0	.165	.731
.2	.291	.779*	.3	.258	.732
.4	.430	.830	.6	.314	.732
.6	.588	.884	1.2	.376	.733
.8	.733	.940	2.0	.410	.733
1.0	1.00*	1.00	3.0	.434	.733
				*	*

* indicates the location of the true value, .75.

References

Harris, D.L. 1964. Expected and predicted progress from index selection involving estimates of population parameters. Biometrics 20:46-72.

Hayes, J.F. and W.G. Hill. 1980. A reparameterization of a genetic selection index to locate its sampling properties. Biometrics 36:237-248.

Hayes, J.F. and W.G. Hill. 1981. Modification of estimates of parameters in the construction of genetic selection indices ('bending'). Biometrics 37:483-493.

Hill, W.G. and R. Thompson. 1978. Probabilities of non-positive definite between-group or genetic covariance matrices. Biometrics 34:429-439.

Hocking, R.R. 1976. The analysis and selection of variables in linear regression. Biometrics 32:1-49.

Hoerl, A.E. and R.W. Kennard. 1970. Ridge regression: biased estimation for non-orthogonal problems. Technometrics 12:55-67.

Sales, J. and W.G. Hill. 1976a. Effect of sampling errors on efficiency of selection indices. 1. Use of information from relatives for single trait improvement. Animal Production 22:1-17.

Sales, J. and W.G. Hill. 1976b. Effect of sampling errors on efficiency of selection indices. 2. Use of information on associated traits for improvement of a single important trait. Animal Production 23:1-14.

Williams, J.S. 1962a. Some statistical properties of a genetic selection index. Biometrika 49:325-337.

Williams, J.S. 1962b. The evaluation of a selection index. Biometrics 18:375-393.