Marker-assisted selection using ridge regression

JOHN C. WHITTAKER¹*, ROBIN THOMPSON² AND MIKE C. DENHAM¹

¹Department of Applied Statistics, University of Reading, PO Box 240, Earley Gate, Reading RG6 2FN, UK
²IACR Rothamsted, Harpenden, Herts AL5 2JQ, UK, and Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS, UK

(Rceived 2 March 1999 and in revised form 29 July 1999)

Summary

In crosses between inbred lines, linear regression can be used to estimate the correlation of markers with a trait of interest; these marker effects then allow marker assisted selection (MAS) for quantitative traits. Usually a subset of markers to include in the model must be selected: no completely satisfactory method of doing this exists. We show that replacing this selection of markers by ridge regression can improve the mean response to selection and reduce the variability of selection response.

1. Introduction

When two inbred lines are crossed, linkage disequilibrium is generated between genetic markers and quantitative trait loci (QTL). Lande & Thompson (1990) noted that this made marker assisted selection (MAS) possible, by using regression of phenotype on markertype to determine the markers associated with the trait because of linkage to QTL and to estimate the correlation between each marker and the trait caused by this linkage. Combination of these marker effects with phenotypic information using a selection index gives a procedure which has been shown by computer simulation to be more effective than selection on phenotype alone when sample sizes are large and heritability low (Zhang & Smith, 1992; Gimelfarb & Lande, 1994a, b; Whittaker et al., 1995). However, Xie & Xu (1998) have pointed out that if the cost of markertyping individuals is allowed for, phenotypic selection may be more cost-effective than MAS at present: widespread adoption of MAS may therefore require further reductions in markertyping costs.

In general, we cannot include all markers in the regression model and so must select a subset of markers to fit. No entirely satisfactory way of doing this exists. Here we evaluate a method which replaces the subset selection procedure by ridge regression, a method which often performs better than subset selection in regression problems when prediction is of primary interest (Breiman, 1995).

2. Methods

We shall consider an F2 population derived from a cross between two inbred lines. We label the alleles at the i\textsuperscript{th} QTL in the first line \(Q_i\), and the alleles at the j\textsuperscript{th} marker locus \(M_j\). The corresponding alleles in the second line are labeled \(q_i\) and \(m_j\). For each individual in the population we know the phenotype \(y\) and the number of \(M_i\) alleles at the \(i\)th marker locus, \(x_i\) so that the marker genotype of an individual is described by \(x = (x_1, x_2, \ldots, x_k)\). From these we wish to construct an estimate \(\hat{z}\) of the genetic value of the individual, \(z\).

Lande & Thompson (1990) suggested using the linear estimator \(\hat{z} = b_0 y + b_1 s\) where, for any individual, the marker score \(s\) is given by

\[
s = \hat{b}_0 + \sum_{i \in A} \hat{b}_i x_i.
\]

Here \(A\) is the set of markers for which effects have been fitted. We have considered the problems involved in calculating \(b_0\) and \(b_1\) elsewhere (Whittaker et al., 1997); in this paper we are concerned with the calculation of the marker score \(s\) and so shall concentrate on the model

\[
\hat{z} = \sum_{i \in A} \hat{b}_i x_i,
\]
where the constant term $\hat{\theta}_0$ has been suppressed for notational convenience. The marker effects $\hat{\theta}_i$ are estimated by the usual least squares estimators

$$\hat{\theta} = (X^tX)^{-1} X^t y,$$

where $y = (y_1, y_2, \ldots, y_n)$ for the $n$ individuals in the population and $X$ is a $n \times k$ matrix whose $i$th row gives the markertype of the $i$th individual in the population.

(i) Subset selection

Usually we cannot include all markers in $\mathcal{A}$. In an extreme case there may be more markers than individuals in the population, but even if this is not so, fitting too many markers increases the variance of the $\hat{\theta}$ and so results in bad estimates of $z$ (Zhang & Smith, 1992). The problem of choosing a linear model so as to trade off the variance and bias of the estimator of interest (here $z$), which increase and decrease respectively as the number of variables included in the model increases, and thus to minimize prediction error, has been much discussed in the statistical literature (Miller, 1990). This remains an active research area.

A number of methods of selecting $\mathcal{A}$ have been suggested. Gimelfarb & Lande (1994a, b) used a forward selection procedure to select the $p$ markers giving the largest reduction in residual sum of squares (RSS). The forward selection procedure works well at selecting the best $p$ markers, in the sense of performing similarly to examination of all possible models whilst being much cheaper computationally, but does not help us choose $p$: rather, $p$ must be chosen in advance by the researcher. As the optimum number of markers to include varies with the number and distribution of QTL, marker map, etc., and is therefore unknown, this seems unsatisfactory. Accordingly Whittaker et al. (1995) developed an automatic procedure based on Mallows’ $C_p$ (Mallows, 1973).

For a linear model with $p$ parameters, Mallows’ $C_p$ is defined as

$$C_p = \frac{RSS}{\hat{\sigma}^2} - n + 2p,$$

where RSS is the residual sum of squares from the model under consideration, $n$ is the number of observations and $\hat{\sigma}^2$ is an estimate of the error variance $\sigma^2$. Choosing $p$ to minimise $C_p$ is a commonly used way of selecting the number of variables to include in a linear model. However, there are well-known problems with this approach. The theoretical justification of $C_p$ relies on comparing a single model with $p$ parameters with another model with $p^*$ parameters. In most applications, including the current one, we compare the best fitting model with $p$ parameters with the best fitting model with $p^*$ parameters. The fact that these models have been selected from a number of possible $p$ and $p^*$ parameter models invalidates this underlying theory and can lead to the selection of far from optimal models. An alternative would be to base a stopping rule for forward selection on $F$-tests, but it is not clear what the appropriate significance threshold is in these subset selection problems (Miller, 1990).

(ii) Ridge regression

An alternative to subset selection is ridge regression (e.g. Myers, 1992). Here all variables are included in the model, but the normal least squares estimators given above are replaced by

$$\hat{\theta} = (X^tX + \lambda I)^{-1} X^t y,$$

where $I$ is the identity matrix. That is, the usual estimates are shrunk towards zero, with the degree of shrinkage determined by the parameter $\lambda$. This shrinking of all estimates towards zero by a constant factor is sensible only if all variables have mean zero and the same variance: therefore it is usual to centre and standardize the variables so that we work with the transformed variables $x_i \rightarrow \frac{x_i - \bar{x}_i}{sd_i}$ where $\bar{x}_i$ and $sd_i$ are the sample mean and standard deviation of the variable $x_i$.

There has been much debate about the advantages and disadvantages of ridge regression (Breiman, 1995), but there is evidence that it outperforms subset selection in some problems. Here we compare the performance of subset selection and ridge regression in MAS by simulation.

To perform ridge regression, we need some way of choosing $\lambda$; a number of ways of doing this exist (Myers, 1992), but we are restricted by our need for an automatic procedure to allow the use of simulation. Write the (unknown) genetic values of the individuals as $z = (z_1, z_2, \ldots, z_n)$ and suppose that for $i = 1, 2, \ldots, n$, $y_i = z_i + \epsilon_i$ with $\epsilon_i$ independent and identically distributed with mean zero and variance $\sigma^2$. We want to choose $\lambda$ to minimize what Breiman (1992) called the model error (ME); here

$$ME = \sum_{i=1}^{n} (z_i - \hat{z}_i)^2,$$

where $\hat{z}_i = \sum_{j=1}^{k} \hat{\theta}_j x_{ij}$ and $\hat{\theta} = (X^tX + \lambda I)^{-1} X^t y$ as above. It can be shown (Mallows, 1973; Breiman, 1992) that a good estimator of the model error is

$$\hat{ME} = RSS - n \sigma^2 + 2 \sigma^2 tr[MX(X^tX + \lambda I)^{-1} X^t],$ $

where RSS is the residual sum of squares from the model under consideration and $tr(M)$ denotes the trace of the $n \times n$ matrix $M$. Thus we can evaluate $\hat{ME}$.
for a range of \( \lambda \) and choose the \( \lambda \) which minimizes \( ME \). In practice, of course, \( \sigma^2_e \) must be replaced by an estimate of the environmental variance, as in the formula for Mallows’ \( C_p \).

Note a further advantage of ridge regression: the addition of the \( \lambda I \) term reduces collinearity and prevents the matrix \( X'X \) becoming singular or near-singular. Singular \( X'X \) often occur after several generations of selection as the population becomes increasingly inbred so this is potentially important here.

3. Simulations

We compared the following three methods: selection based solely on an individual’s phenotypic value (PHENO), selection based on \( \hat{z} = \Sigma_{i=1}^s \hat{a_i} x_i \), where \( a_i \) is chosen using the Mallow’s \( C_p \) scheme described in Whittaker et al. (1995) (MARKER) and selection based on ridge regression (RIDGE) as outlined above.

In all simulations, QTL are assumed to combine in a multiplicative manner and are increasingly inbred so this is potentially important here.

The results obtained are shown in Table 1 as percentages of the maximum genetic value obtainable, that is the genetic value of an individual possessing all favourable alleles. As usual, marker-assisted methods are increasingly favoured, in comparison with selection on phenotype, by increasing population size and decreasing heritability (e.g. Moreau et al., 1998; Van Berloo & Stam, 1999). In all cases RIDGE performs slightly better than both MARKER and PHENO. Also, the standard errors of selection response are smaller for RIDGE than for MARKER, so that RIDGE is the more reliable selection method. This is probably because ridge regression is a very stable procedure in the sense that small changes in the data do not produce large changes in the estimated regression coefficients: subset selection is unstable and so produces more variable response to selection.

Table 1. Accumulated selection response (standard error) after 1, 2, 4, 8 generations of selection, for (a) \( n = 100, h^2 = 0.1 \); (b) \( n = 100, h^2 = 0.2 \); (c) \( n = 400, h^2 = 0.1 \) and (d) \( n = 400, h^2 = 0.2 \)

<table>
<thead>
<tr>
<th>Generation</th>
<th>PHENO</th>
<th>MARKER</th>
<th>RIDGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) 1</td>
<td>0.0721 (0.0181)</td>
<td>0.1114 (0.0241)</td>
<td>0.1268 (0.0214)</td>
</tr>
<tr>
<td>2</td>
<td>0.1364 (0.0253)</td>
<td>0.2056 (0.0340)</td>
<td>0.2228 (0.0287)</td>
</tr>
<tr>
<td>4</td>
<td>0.2525 (0.0342)</td>
<td>0.3289 (0.0477)</td>
<td>0.3561 (0.0402)</td>
</tr>
<tr>
<td>8</td>
<td>0.4166 (0.0412)</td>
<td>0.4681 (0.0588)</td>
<td>0.5024 (0.0474)</td>
</tr>
<tr>
<td>(b) 1</td>
<td>0.1033 (0.0236)</td>
<td>0.1413 (0.0278)</td>
<td>0.1528 (0.0250)</td>
</tr>
<tr>
<td>2</td>
<td>0.1944 (0.0330)</td>
<td>0.2503 (0.0386)</td>
<td>0.2663 (0.0337)</td>
</tr>
<tr>
<td>4</td>
<td>0.3349 (0.0397)</td>
<td>0.3907 (0.0449)</td>
<td>0.4185 (0.0354)</td>
</tr>
<tr>
<td>8</td>
<td>0.5037 (0.0362)</td>
<td>0.5348 (0.0440)</td>
<td>0.5669 (0.0354)</td>
</tr>
<tr>
<td>(c) 1</td>
<td>0.0745 (0.0128)</td>
<td>0.1621 (0.0162)</td>
<td>0.1650 (0.0151)</td>
</tr>
<tr>
<td>2</td>
<td>0.1435 (0.0181)</td>
<td>0.2686 (0.0227)</td>
<td>0.2749 (0.0185)</td>
</tr>
<tr>
<td>4</td>
<td>0.2589 (0.0215)</td>
<td>0.3945 (0.0283)</td>
<td>0.4182 (0.0233)</td>
</tr>
<tr>
<td>8</td>
<td>0.4301 (0.0218)</td>
<td>0.5112 (0.0309)</td>
<td>0.5445 (0.0244)</td>
</tr>
<tr>
<td>(d) 1</td>
<td>0.1031 (0.0128)</td>
<td>0.1810 (0.0162)</td>
<td>0.1806 (0.0151)</td>
</tr>
<tr>
<td>2</td>
<td>0.1944 (0.0181)</td>
<td>0.2923 (0.0227)</td>
<td>0.2979 (0.0185)</td>
</tr>
<tr>
<td>4</td>
<td>0.3375 (0.0215)</td>
<td>0.4319 (0.0283)</td>
<td>0.4537 (0.0233)</td>
</tr>
<tr>
<td>8</td>
<td>0.5094 (0.0218)</td>
<td>0.5537 (0.0309)</td>
<td>0.5787 (0.0244)</td>
</tr>
</tbody>
</table>
would be interesting to examine their performance in MAS. The Bayesian interpretation of ridge regression also suggests a number of promising alternative methods. In addition, reversible jump MCMC (Green, 1995; Sillanpää & Arjas, 1998) is another approach that has great potential.

References


