STATISTICAL PROBLEMS IN MARKER ASSISTED SELECTION FOR QTL

Rohan L. Fernando
Department of Animal Sciences, University of Illinois,
1207 W. Gregory Dr. Urbana, Il 61801, USA

SUMMARY

Selection of animals based on the conditional mean of the genotype given all data maximizes expected genetic progress. The genotypic model used for marker assisted selection leads to a conditional distribution that is the weighted sum of normal densities. The conditional mean is the weighted sum of conditional means, and the likelihood of the data is the sum of weighted normal likelihoods. In all of the above, the sum is over all possible k^n marked QTL genotypes, where k is the number of genotypes at the marked QTL, and n the number of animals. In human genetics, computable algorithms for the likelihood have been devised by writing it as the product of conditional densities. Such algorithms are not available for the conditional mean. Approximations for the conditional mean and likelihood are discussed.

INTRODUCTION

Selection theory for quantitative traits has been based, almost exclusively, on an additive genetic model. The genotypic effect in this model is assumed to be the sum of effects of genes at a large number of loci, with each gene having a small effect. Such loci are called quantitative trait loci (QTL).

In traditional selection procedures, phenotypic data are used to predict additive genetic values. Animals with the highest predicted values are then selected for breeding purposes. Phenotypic data do not provide information on individual QTL genes. However, with the advances in biochemical and molecular genetics, information on QTL genes is becoming available in the form of genetic marker data. Such data consists of genotypic information at genetically polymorphic loci that are closely linked to QTL. This paper addresses the statistical problems in the use of marker data together with phenotypic data for selection in outbred populations.

NOTATION AND ASSUMPTIONS

Marker assisted selection will be considered with marker data at one locus. Let M_i^p (M_i^m) be the marker allele that animal i inherited from its paternal parent (maternal parent) The QTL locus linked to the marker locus will be referred to as the marked QTL (MQTL). Let Q_i^p (Q_i^m) be the MQTL allele animal i inherited from its paternal parent (maternal parent). In outbred populations there will not be a consistent association between the marker genotype and the MQTL genotype, unless there is linkage disequilibrium. Therefore, in an outbred population, knowing the marker genotype of an animal does not provide information about its MQTL genotype. However, for an animal i with sire s, knowing which of the sire's marker alleles, M_s^p or M_s^m, was inherited by animal i, provides useful information. For example, if animal i receives a copy of M_s^m from its sire, then the probability that Q_i^p is a copy of Q_s^p is r and that it is a copy of Q_s^m is $(1-r)$, where r is the frequency of recombination between the marker locus and the MQTL. Let $\boldsymbol{M}$ be this type of marker information, and $\boldsymbol{y}$ be phenotypic information.

The additive model

$$a_i = v_i^p + v_i^m + u_i \tag{1}$$

will be used for the genotypic value (a_i) of animal i, where v_i^p and v_i^m are the effects of MQTL alleles Q_i^p and Q_i^m, respectively, and u_i is the additive effect of the remaining QTL alleles. The phenotypic value of animal i will be modeled as $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + a_i + e_i$, where $\mathrm{E}(y_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$ and e_i is a residual. It is assumed that u_i and e_i are normally distributed.

PREDICTION AND ESTIMATION

When a constant number of animals is to be selected, it has been shown that selection based on the conditional mean of the additive genetic value given all the data maximizes expected genetic progress (Goffinet, 1983; Fernando and Gianola 1986). Thus, in marker assisted selection, we would like to predict $\boldsymbol{a}$ using $\mathrm{E}(\boldsymbol{a}|\boldsymbol{y},\boldsymbol{M})$.

The conditional distribution of $\boldsymbol{a}$ given $\boldsymbol{y}$ and $\boldsymbol{M}$ can be written as

$$f(\boldsymbol{a} \mid \boldsymbol{y}, \boldsymbol{M}) = \sum_{\boldsymbol{Q}} f(\boldsymbol{a} \mid \boldsymbol{y}, \boldsymbol{Q}) \Pr(\boldsymbol{Q} \mid \boldsymbol{y}, \boldsymbol{M}) \tag{2}$$

where $\boldsymbol{Q}$ is a $n \times 2$ matrix with row i containing Q_i^p and Q_i^m. The summation in (2) is over all possible values of $\boldsymbol{Q}$. Under the assumptions made here, $f(\boldsymbol{a} \mid \boldsymbol{y}, \boldsymbol{Q})$ is a normal density. The probability in (2) can be written as

$$\Pr(\boldsymbol{Q} \mid \boldsymbol{M}, \boldsymbol{y}) \propto f(\boldsymbol{y} \mid \boldsymbol{Q}, \boldsymbol{M}) \Pr(\boldsymbol{Q} \mid \boldsymbol{M}) \tag{3}$$

The first term in (3) is a normal density. Following the strategy used in human genetics (e.g., Elston and Stewart, 1971; Bonney, 1984), the second term can be written as

$$\begin{aligned} \Pr(\boldsymbol{Q} \mid \boldsymbol{M}) &= \Pr(Q_1^p \mid \boldsymbol{M}) \Pr(Q_1^m \mid Q_1^p, \boldsymbol{M}) \cdots \\ &\quad \Pr(Q_n^m \mid Q_1^p, Q_1^m, \ldots, Q_{n-1}^p, Q_{n-1}^m, Q_n^p, \boldsymbol{M}) \\ &= \prod_{i=1}^{n} \pi_i^p \pi_i^m \end{aligned} \tag{4}$$

where $\pi_i^p = \Pr(Q_i^p)$, the frequency of Q_i^p in the population, if the sire (s) of animal i is not in the pedigree; or $\pi_i^p = \Pr(Q_i^p \mid Q_s^p, Q_s^m, \boldsymbol{M})$, if s is in the pedigree. Similarly, π_i^m is defined by conditioning on the dam's paternal and maternal MQTL alleles, Q_d^p and Q_d^m. These conditional probabilities depend on marker information only if the parent is heterozygous. For example, if the sire (s) of animal i has MQTL alleles Q_s^p=A and Q_s^m=a, the probability of Q_i^p being equal to A or a is given in table 1. If s has MQTL genotype AA, $\Pr(Q_i^p$=A)=1 and $\Pr(Q_i^p$=a)=0 regardless of the marker data. Similarly, if s has MQTL genotype aa, $\Pr(Q_i^p$=A)=0 and $\Pr(Q_i^p$=a)=1. The conditional probabilities for Q_i^m are similarly defined.

If the $\mathrm{E}(\boldsymbol{y})$, and second moments of $\boldsymbol{a}$ and $\boldsymbol{y}$, given $\boldsymbol{Q}$, are known, $\mathrm{E}(\boldsymbol{a}|\boldsymbol{y},\boldsymbol{Q})$ can be calculated as a linear function of $\boldsymbol{y}$, and $\mathrm{E}(\boldsymbol{a}|\boldsymbol{y},\boldsymbol{M})$ can be written as

$$\mathrm{E}(\boldsymbol{a} \mid \boldsymbol{y}, \boldsymbol{M}) = \sum_{\boldsymbol{Q}} \mathrm{E}(\boldsymbol{a} \mid \boldsymbol{y}, \boldsymbol{Q}) \Pr(\boldsymbol{Q} \mid \boldsymbol{y}, \boldsymbol{M}) \tag{5}$$

Table 1: Probability that Q_i^p is A or a given its sire has MQTL genotype Aa and $M_i^p = M_s^p$ or $M_i^p = M_s^m$

Marker Data	Probability	
	Q_i^p=A	Q_i^p=a
$M_i^p = M_s^p$	1-r	r
$M_i^p = M_s^m$	r	1-r

If the $E(\boldsymbol{y})$ is not known, $E(\boldsymbol{a}|\boldsymbol{y},\boldsymbol{Q})$ cannot be calculated, and thus $E(\boldsymbol{a}|\boldsymbol{y},\boldsymbol{M})$ cannot be calculated. However, for $\boldsymbol{w}$, a vector of records "corrected" for $E(\boldsymbol{y})$, $E(\boldsymbol{a}|\boldsymbol{w},\boldsymbol{M})$ can be written as

$$E(\boldsymbol{a} \mid \boldsymbol{w}, \boldsymbol{M}) = \sum_{\boldsymbol{Q}} E(\boldsymbol{a} \mid \boldsymbol{w}, \boldsymbol{Q}) \Pr(\boldsymbol{Q} \mid \boldsymbol{w}, \boldsymbol{M}) \tag{6}$$

The conditional mean, $E(\boldsymbol{a}|\boldsymbol{w},\boldsymbol{Q})$, in (6) is the best linear unbiased predictor (BLUP) of $\boldsymbol{a}$ given $\boldsymbol{Q}$ (Goffinet, 1983; Fernando and Gianola, 1986), and can be calculated using Henderson's mixed model equations (Henderson, 1973). The probability in (6) is defined in the same way as that in (2), given by (3), except with $\boldsymbol{w}$ in place of $\boldsymbol{y}$.

These calculations require knowledge of the population frequencies and effects of the MQTL alleles, the recombination rate r, and the variances of the u_i's and the e_i's. These parameters can be estimated by maximum likelihood. The likelihood to be maximized is

$$L(\boldsymbol{y}) = \sum_{\boldsymbol{Q}} f(\boldsymbol{y} \mid \boldsymbol{Q}, \boldsymbol{M}) \Pr(\boldsymbol{Q} \mid \boldsymbol{M}) \tag{7}$$

As stated earlier, $f(\boldsymbol{y} \mid \boldsymbol{Q}, \boldsymbol{M})$ is a normal density.

The problem in working with (5), (6), or (7) is that, even with two MQTL alleles segregating in the population, the summation $\sum_{\boldsymbol{Q}}$ is over 2^{2n} values that $\boldsymbol{Q}$ can take. This problem has been addressed in human genetics in the context of estimation and hypothesis testing by maximum likelihood (Elston, 1990).

Elston and Stewart (1971) outlined how the likelihood of "simple" pedigree data could be calculated when a_i is either discrete or continuous. A simple pedigree cannot have the maternal and paternal grand parents of an individual in the pedigree, nor can it have "loops" (Ott, 1974). Lange and Elston (1974) have given an exact definition for a simple pedigree, and have described algorithms to accommodate more complex ones.

When the genotypic value is the sum of discrete and continuous components, as in (1), fast algorithms for exact calculation of the likelihood are not available (Elston, 1990). Approximations to the likelihood have been presented by Morton and MacLean (1974) and Hasstedt (1982). The method of Hasstedt was applied to a pedigree with 163 members (Hasstedt, 1982). This would be impossible using expression (7), because the likelihood is a weighted sum of of 2^{326} normal likelihoods.

The algorithms used in human genetics for computing likelihoods are based on writing the likelihood as a product of conditional densities. Although this can always be done, if the pedigree is complex and large, it may still not lead to a computable expression for (7). Further, this strategy does not provide computable expressions for (5) or (6).

An alternative strategy that has been suggested is to approximate (7) by taking the sum over a subset of the values that $\boldsymbol{Q}$ can take. This subset may be chosen at random or in some systematic manner. This approach is analogous to Monte Carlo integration (Bauwens, 1984), and has the advantage that it can also be used to approximate (5) and (6). The accuracy of this approximation needs to be examined.

Yet another alternative is to predict $\boldsymbol{a}$ by BLUP. In this approach BLUP's of the v_i's and u_i's are obtained using Henderson's mixed model equations, and BLUP of a_i is obtained as

$$\hat{a}_i = \hat{v}_i^p + \hat{v}_i^m + \hat{u}_i$$

where $\hat{v}_i^p$, $\hat{v}_i^m$ and $\hat{u}_i$ are the BLUP's of v_i^p, v_i^m and u_i, respectively. The inverse of the variance covariance matrix of the MQTL effects is required for obtaining BLUP through the mixed model equations. Fernando and Grossman (1989) have given an algorithm for obtaining this inverse in linear time. The order of the mixed model equations used here is $p + 3n$, where p is the order of $\boldsymbol{\beta}$. They can be solved using iterative techniques currently employed to solve the usual mixed model equations (Misztal and Gianola, 1987).

The BLUP of $\boldsymbol{a}$ will be a good approximation for (6), if the MQTL effects are approximately normally distributed. Further, if the MQTL effects are normal, $\boldsymbol{y}$ has a normal likelihood, and estimation of genetic parameters by maximum likelihood with relatively large complex pedigrees is feasible (Weller and Fernando, 1990).

REFERENCES

BAUWENS, L. 1984. *Bayesian Full Information Analysis of Simultaneous Equations Models Using Integration by Monte Carlo.* Springer-Verlag, Germany.

BONNEY, G. E. 1984. *Amer. J. Medical Genet.*, 18:731–749.

ELSTON, R. C. 1990. Models for discrimination between alternative modes of inheritance. In GIANOLA, D. and HAMMOND, K. editors, *Advances in Statistical Methods for Genetic Improvement of Livestock*, Springer-Verlag, (in press).

ELSTON, R. C. STEWART, J. 1971. *Human Hered.*, 21:523–542.

FERNANDO, R. L. and GROSSMAN, M. 1989. *Genet. Sel. Evol.* (In press)

FERNANDO, R. L. and GIANOLA, D. 1986. *Theor. Appl. Genet.*, 72:822–825.

GOFFINET, B. 1983. *Genet. Sel. Evol.*, 15:91–98.

HASSTEDT, S. J. 1982. *Comput. Biomed. Res.*, 15:295–307.

HENDERSON, C. R. 1973. In *Anim. Breed. Genet. Symp. in Honor of Dr. J. L. Lush*, pages 10–41, Amer. Soc. Anim. Sci. and Amer. Dairy Sci. Assoc. Champaign, IL.

LANGE K. and ELSTON, R. C. 1974. *Human Hered.*, 25:95–105.

MISZTAL, I, GIANOLA, D. *J. Dairy Sci.*, 1987, 70:716–723.

MORTON, N. E. and MACLEAN, C. J. 1974. *Amer. J. Human Genet.*, 26:489–503.

OTT, J. 1974. *Amer. J. Human Genet.*, 26:588–597.

WELLER, J. and FERNANDO, R. L. 1990. Strategies for the improvement of animal production using marker assisted selection. In *Gene Mapping: Strategies, Techniques and Applications.* Marcel Dekker, Inc. (submitted)