ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS DE MATHEMATICA
Volume 8, 2004

About the data designs for estimation of genetic parameters in animal breeding studies

Tanel Kaart

Abstract. The simple random one-way linear model and ANOVA estimators of random effects, variance components and intraclass correlation are studied. A related sire model in animal genetic studies is considered. The study is focused on the effect of data design on the accuracy of estimates. The patterns of sampling variances of estimates depending on the genetic determination of the trait and data design are examined and optimal designs for different genetic parameters are compared in balanced case.

1. Introduction

Estimation of genetic determination of economically important traits in animal breeding has more than 80 years been based on the comparison of progeny records of different sires. Mixed linear model with random sire effect and fixed environmental effects, known as the sire model, is the main tool implemented for this. Nowadays, due to the great progress in data collection, designs and models in last decades, the simple sire model is applied only in the pilot studies about previously not examined traits. As these studies are usually small, the estimates are not very accurate. But as these studies are also carefully planned, the maximum possible accuracy of estimates is guaranteed by appropriate data design.

Being aware of this, it is surprising that even for the simplest models the number of articles studying data designs for estimation of different population genetic parameters is quite limited and can be dated back to the 1960s and 1970s. The accuracy of estimates alongside the optimal designs is not examined, as a rule.

Received February 10, 2004.

2000 *Mathematics Subject Classification.* 62K05, 62J10, 92D10, 92D50.

Key words and phrases. Optimal design, analysis of variance, genetic parameters, estimation.

In this article basic genetic parameters usually estimated from sire model are studied. These are the effects of sires, interpreted as half of sires additive genetic values (breeding values) or transmitting abilities of sires, the variance component associated with sire effects and the heritability coefficient. The last one is calculated as four times the intraclass correlation (ratio of sires variance component to the total variance). The mean square errors of best linear predictions of sire effects, the variance of estimates of variance components and heritability and the probability to get the heritability estimates that are outside the parameter space, i.e. negative or greater than one, are examined. The formulas and propositions collected from different resources or derived by the author are presented without proofs. The patterns of examined accuracy parameters are found by simulation studies and optimum designs are calculated by exact formulas or by simulations.

Despite that the study is focused only on balanced designs and on estimates obtained by analysis of variance, the tendencies apply also to the unbalanced case and to other variance components estimation methods, as shortly indicated in the last section.

2. Mathematical framework

2.1. Model and estimates. Consider the mixed linear model

$$y_{ij} = \mu + u_i + e_{ij}\,, \tag{1}$$

or in matrix notation

$$\mathbf{y} = \mathbf{1}_N\mu + \mathbf{Z}\mathbf{u} + \mathbf{e}\,,$$

where $\mathbf{y}$ is the $N \times 1$ vector of observed values, μ is the only fixed effect in the model (mean), $\mathbf{1}_N' = (\ 1 \ \ \dots \ \ 1\ )_N'$ and $\mathbf{Z} = \mathbf{I}_a \otimes \mathbf{1}_n$ are known design matrices of order $N \times 1$ and $N \times a$ respectively, associating fixed and random effects with $\mathbf{y}$, $\mathbf{u}' = (\ u_1 \ \ \dots \ \ u_a\ )'$ is a vector of random sire effects, $\mathbf{e}$ is a $N \times 1$ vector of random residuals, and traditionally the number of levels in random factor (number of sires) is marked as a and the number of objects (daughters) per level (sire) i is n in the one way model.

The expectation and the variance-covariance structure are represented as

$$\mathrm{E}(\mathbf{y}) = \mu\,, \quad \mathrm{Var}(\mathbf{u}) = \sigma_u^2\mathbf{I}_a\,, \quad \mathrm{Var}(\mathbf{e}) = \sigma_e^2\mathbf{I}_N\,, \quad \mathrm{Cov}(\mathbf{u},\mathbf{e}') = \mathbf{0}$$

and

$$\mathrm{Var}(\mathbf{y}) = \mathbf{I}_a \otimes (\sigma_u^2\mathbf{J}_n + \sigma_e^2\mathbf{I}_n)\,.$$

Also it holds that

$$\frac{MS(u)}{n\sigma_u^2 + \sigma_e^2}\Big/\frac{MS(e)}{\sigma_e^2} \sim F_{a-1,N-a}\,, \tag{2}$$

where $MS(u)$ and $MS(e)$ indicate the mean squares from the ANOVA table.

If we assume variance components σ_u^2 and σ_e^2 known, the best linear unbiased prediction (BLUP) of $\mathbf{u}$ is

$$\hat{\mathbf{u}} = \sigma_u^2 \mathbf{Z}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{1}_N \hat{\mu}),$$

or component-wise

$$\hat{u}_i = \frac{\sigma_u^2}{\sigma_e^2 + n\sigma_u^2} \sum_{j=1}^{n} (y_{ij} - \hat{\mu}), \text{where } \hat{\mu} = \sum_{i=1}^{N} y_i \Big/ N.$$

The ANOVA estimators of variance components σ_u^2 and σ_e^2 are obtained by equating the mean squares with their expected values and are expressed as

$$\hat{\sigma}_u^2 = \frac{1}{n}[MS(u) - MS(e)]$$

and

$$\hat{\sigma}_e^2 = MS(e).$$

In genetic studies the intraclass correlation coefficient, which measures the magnitude of random genetic effects, is calculated as the ratio of variances, and estimated additionally:

$$\hat{\rho} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2} = \frac{MS(u) - MS(e)}{MS(u) + (n-1)MS(e)}. \tag{3}$$

In sire model 1 the intraclass correlation equals with one quarter of heritability:

$$h^2 = \frac{4\sigma_u^2}{\sigma_u^2 + \sigma_e^2} = 4\rho.$$

2.2. Accuracy of estimates. The mean square errors of BLUP($\mathbf{u}$) and BLUP(u_i) are

$$\text{MSE}(\hat{\mathbf{u}}) = \frac{\sigma_u^2(a\sigma_e^2 + n\sigma_u^2)}{\sigma_e^2 + n\sigma_u^2}$$

and

$$\text{MSE}(\hat{u}_i) = \frac{\sigma_u^2(a\sigma_e^2 + n\sigma_u^2)}{a(\sigma_e^2 + n\sigma_u^2)}, \tag{4}$$

respectively, and the sampling variance of $\hat{\sigma}_u^2$ is

$$\text{Var}(\hat{\sigma}_u^2) = \frac{2}{n^2}\left[\frac{(n\sigma_u^2 + \sigma_e^2)^2}{a-1} + \frac{\sigma_e^4}{a(n-1)}\right].$$

All these results are well known and can be found in many textbooks (Searle et al, 1992; Khuri et al, 1998).

For the sampling variance of the estimator of the intraclass correlation coefficient there does not exist exact formula even in the balanced case. Usually an approximate formula is used:

$$\mathrm{Var}(\hat{\rho}) \approx \frac{2[1+(n-1)\rho]^2(1-\rho)^2}{n(n-1)(a-1)}\,, \tag{5}$$

derived by Osborne and Paterson (1952) using a first-order Taylor-series expansion of equality 3. Zerbe and Goldgar (1980) derived an alternative formula based on the F-ratio 2. Their derivation is based on the approximation

$$\mathrm{Var}[f(w)] \approx [\partial f(w)/\partial w]^2\,\mathrm{Var}(w)\,,$$

which gave the following final formula:

$$\mathrm{Var}(\hat{\rho}) \approx [1+(n-1)\rho]^2(1-\rho^2)\frac{2[a(n-1)]^2(an-3)}{n^2(a-1)[a(n-1)-2]^2[a(n-1)-4]}\,. \tag{6}$$

Visscher (1998) examined different expressions of the sampling variances of intraclass correlations and concluded that formula 5 gives quite precise estimate to the sampling variance of $\hat{\rho}$, except for a small number of sires and/or a large heritability coefficient. Then the sampling variance was underestimated. As $\mathrm{Var}(\hat{\rho})$ calculated by 6 is always bigger than the corresponding value based on 5, then for large heritability values and/or small number of sires the approximation of Zerbe and Goldgar seems to be appropriate. In the following analysis only Osborne, Paterson formula 5 is applied because the optimal designs are the same for both approximations of $\mathrm{Var}(\hat{\rho})$.

A supplementary undesirable property of ANOVA estimates of variance components is that the estimates can fall outside the parameter space. The probability of negative variance component estimate (which is equivalent to the negative heritability estimate) is reviewed in Searle et al (1992) and Khuri et al (1998) and is expressed as

$$\mathrm{P}(\hat{h}^2 < 0) = \mathrm{P}(\mathrm{F}_{a(n-1),\,a-1} > 1 + n\tau)\,,$$

where $\tau = \sigma_u^2/\sigma_e^2$ and $\mathrm{F}_{(a-1),a(n-1)}$ denotes random variable with F-distribution with a-1 and $a(n$-1) degrees of freedom. Similarly the probability to get the heritability estimate from sire model bigger than one is derived by the author (Kaart, 1997):

$$\mathrm{P}(\hat{h}^2 > 1) = \mathrm{P}(\mathrm{F}_{a(n-1),\,a-1} < (n/3+1)^{-1}(1+n\tau))\,.$$

2.3. Optimal designs. It is natural to suppose for balanced data that both, number of sires a and number of daughters per sire n are bigger than one. Then it is obvious that $\mathrm{MSE}(\hat{\mathbf{u}})$ is minimized when the number of groups is minimal, $a = 2$, and the number of observations per group is maximal, $n = N/2$.

The optimal number of daughters per sire minimizing $\mathrm{MSE}(\hat{u}_i)$ is found by the author considering n as a continuous argument and studying the derivatives of equation 4, after what it is seen that

$$n = \frac{-1+\sqrt{1+N\tau}}{\tau}. \tag{7}$$

In a similar way the number of daughters per sire n which minimizes $\mathrm{Var}(\hat{\sigma}_u^2)$ has been derived already by Hammersley in 1948 (Hammersley, 1948; Khuri, 2000):

$$n = \frac{N(\tau+1)+1}{N\tau+2}. \tag{8}$$

It can be shown that the design optimal for $\mathrm{Var}(\hat{\sigma}_u^2)$ is also optimal for $\mathrm{Var}(\hat{\rho})$.

To illustrate the derived criteria of optimality, to find out the optimal data design at the point of inadmissible heritability estimates and to examine the accuracy of estimates besides the optimal designs, the patterns of $\mathrm{MSE}(\hat{u}_i)$, $\mathrm{Var}(\hat{\sigma}_u^2)$, $\mathrm{Var}(\hat{h}^2)$ and $\mathrm{P}(\hat{h}^2 < 0) + \mathrm{P}(\hat{h}^2 > 1)$ were found and both integer and continuous optimum numbers of daughters per sire were calculated by formulas 7 and 8 or by simulations. Without loss of generality the error variance was taken equal to one. The patterns were drawn for the data size $N = 360$, a reasonable number for small practical experiments and also giving a possibility to be divide data into groups with equal integer size in different ways.

3. Results

In Figure 1 the pattern of $\mathrm{MSE}(\hat{u}_i)$, in Figure 2 the pattern of $\mathrm{Var}(\hat{\sigma}_u^2)$, in Figure 3 the expanded pattern of $\mathrm{Var}(\hat{h}^2)$ and in Figure 4 the pattern of $\mathrm{P}(\hat{h}^2 < 0) + \mathrm{P}(\hat{h}^2 > 1)$ are presented.

For all studied parameters the accuracy of estimates is stronger influenced by the number of sires than the number of daughters per sire. The precise estimation of variance components and their functions (heritability) requires bigger number of groups compared to the precise prediction of realised values of random effects. In all cases the deficiency of sires increases the inaccuracy of estimates when the effect of sires, measured via heritability, is large. In estimating variance components a small number of sires may cause dramatic loss of accuracy (Figures 2 and 3). For small heritability values the estimates are more accurate and do not depend so much on the design. As the heritability of trait increases the accuracy of estimates decrease and the optimum number of sires increase. The probability of heritability estimates negative or greater than one depends not so much on the number of sires than the real population value of heritability – if the real parameter value is close to zero then it is obvious that the probability of getting negative estimate increases.

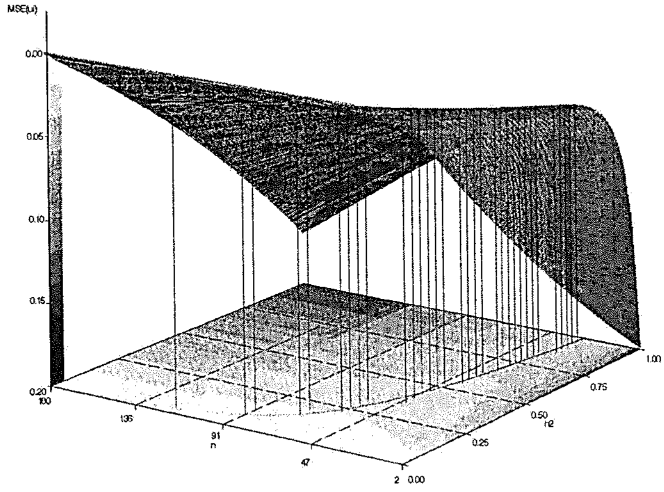

FIGURE 1. Pattern of MSE($\hat{u}_i$) and optimal number of daughters per sire (vertical arrows for integers and dotted line for real numbers) in different true heritability values when $N = 360$ and $\sigma_e^2 = 1$.

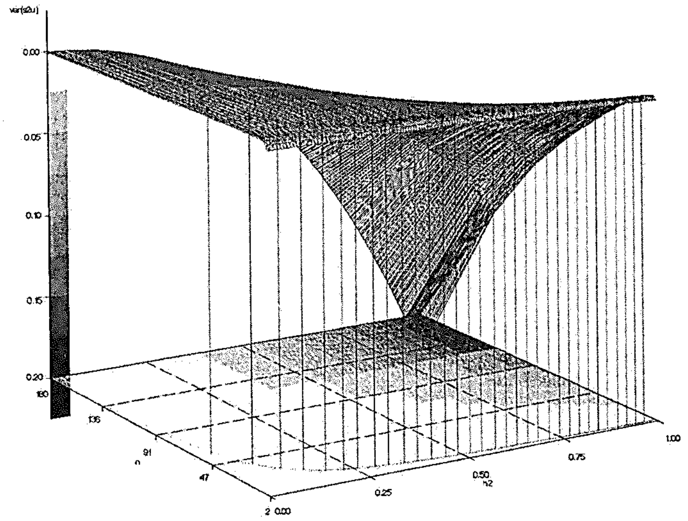

FIGURE 2. Pattern of Var($\hat{\sigma}_u^2$) and optimal number of daughters per sire (vertical arrows for integers and dotted line for continuous numbers) in different true heritability values when $N = 360$ and $\sigma_e^2 = 1$.

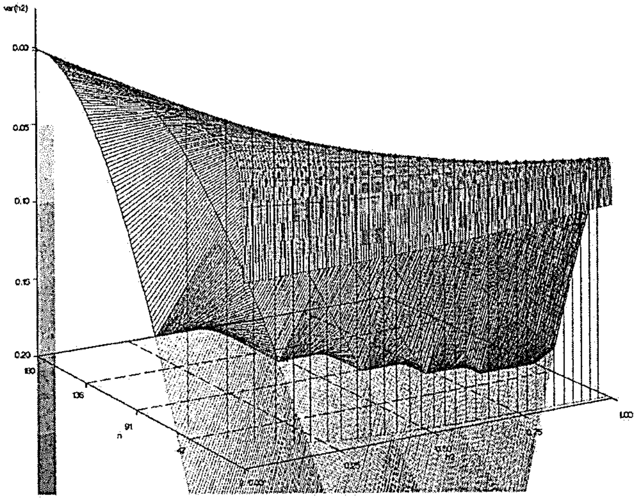

FIGURE 3. Pattern of $\text{Var}(\hat{h}^2)$ and optimal number of daughters per sire (vertical arrows for integers and dotted line for real numbers) in different true heritability values when $N = 360$ and $\sigma_e^2 = 1$.

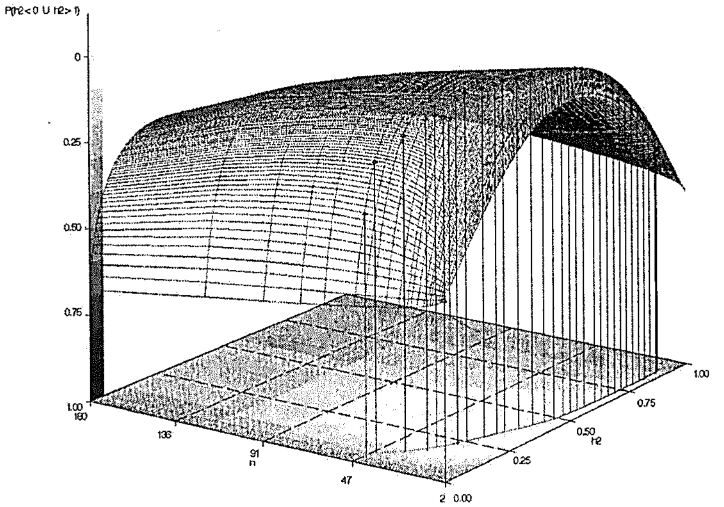

FIGURE 4. Pattern of probabilities to get heritability estimates negative or greater than one and optimal number of daughters per sire in different true heritability values when $N = 360$ and $\sigma_e^2 = 1$.

4. Discussion and conclusions

The considered tendencies apply also in much bigger datasets. For example, the probability to get heritability estimates that are outside the parameter space, come close to zero everywhere except the regions near to zero or one.

It appears, that the usual assumption of animal breeders to increase the number of daughters per sire for increasing the accuracy of estimates is not the best way. Surely, increasing number of daughters per sire will increase the accuracy of estimates, but increasing the number of sires has a bigger effect. This misconception is shortly also noted by Searle et al (1992, p 68-69). Studies of unbalanced designs (not presented here) indicated that in some cases even adding small number of sires with only one daughter will decrease the variability of estimates.

Khuri et al (1998, p 56-61) have shown that Var($\hat{\sigma}_u^2$) and P($\hat{\sigma}_u^2 < 0$) is minimized when the dataset is balanced compared with unbalanced designs. The same effect can be shown also for MSE($\hat{u}_i$), for Var($\hat{h}^2$) and for P($\hat{h}^2 >$ 1).

For unbalanced datasets the criteria of optimal design are not clear. Estimation of variance components studies (which can be relied on the estimation of heritability coefficient) concluded that the design with the closest number of classes (sires) should be used to get the optimum (Anderson, Crump, 1967). Norell (2001) showed that this suggestion does not always yield the minimum of Var($\hat{\sigma}_u^2$). Moreover, Shen et al (1996) studied maximum likelihood estimates of amount of quantitative genetic parameters expressed as ratio of functions of variance components and recognized that for some functions the balanced designs do not always give the most efficient estimates.

In this article it was assumed for MSE($\hat{u}_i$) that variance components (or their ratio or heritability) were already known. This is the common assumption when the animals are ranked by their genetic values and was also suggested by Koots and Gibson (1996), who examined different studies and concluded that the variability of heritability estimates is much bigger than could be expected by the theory. If the genetic determination of studied trait is not known, the variance components need to be estimated first, after what these estimates are used to calculate so-called second stage predictors or estimated BLUPs (EBLUP) of random effects. Then, as shown by Kackar, Harville (1984) and Das et al (2004), the expression of MSE($\hat{u}_i$) contains additional term depending on the sampling variance of estimates of variance components. The studies about data designs in this situation are yet unavailable.

References

Anderson, R. L. and Crump, P. P. (1967). Comparison of designs and estimaton procedures for estimating parameters in a two-stage nested process. *Technometrics* **9**, 499–516.

Das, K., Jiang, J. and Rao, J. N. K. (2004). Mean squared error of empirical predictor. *Ann. Statist.* **32**, 818–840.

Hammarsley, J. M. (1949). The unbiased estimate and standard error of the interclass variance. *Metron* **15**, 189–205.

Kaart, T. (1997). Probability of the estimation of heritability being negative or greater than one. *In: Proceedings of the 3rd Baltic Animal Breeding Conference*, Latvian Ministry of Agriculture, Department of Agriculture Strategy and Co-operation, Riga, 57–59.

Kackar, R. N. and Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *J. Amer. Statist. Assoc.* **79**, 853–862.

Khuri, A. I. (2000). Designs for variance components estimation: past and present. *Internat. Statist. Rev.* **68**, 311–322.

Khuri, A. I., Mathew, T. and Sinha, B. K. (1998). *Statistical Tests for Mixed Linear Models*. Wiley, New York.

Koots, K. R. and Gibson, J. P. (1996). Realized sampling variances of estimates of genetic parameters and the difference between genetic and phenotypic correlations. *Genetics* **143**, 1409–1416.

Norell, L. (2001). *On the Optimum Number of Levels in the One-way Model With Random Effects*. Rapport 65, Department of Biometry and Informatics, Swedish University of Agricultural Sciences, Uppsala.

Osborne, R. and Paterson, W. S. B. (1952). On the sampling variance of heritability estimates derived from variance analysis. *Proc. Roy. Soc. Edinburgh Sect. B* **64**, 456–461.

Searle, S. R., Casella, G. and McCulloch, C. E. (1992). *Variance Components*. Wiley, New York.

Shen, P.-S., Cornelius, P. L. and Anderson, R. L. (1996). Planned unbalanced designs for estimation of quantitative genetic parameters. I: Two-way matings. *Biometrics* **52**, 56–70.

Visscher, P. M. (1998). On the sampling variance of intraclass correlations and genetic correlations. *Genetics* **149**, 1605–1614.

Zerbe, G. O. and Goldgar, D. E. (1980). Comparison of intraclass correlation coefficients with the ratio of two independent F-statistics. *Comm. Statist. Theory Methods* Ser. A **9**, 1641–1655.

Estonian Agricultural University, Institute of Animal Science, Kreutzwaldi 1, 51014 Tartu, Estonia

E-mail address: `ktanel@eau.ee`