Genet. Sel. Evol. 30 (1998) 565–583

Original article

Attenuating effects of preferential treatment with Student-t mixed linear models: a simulation study

Ismo Strandén [a,b*], Daniel Gianola [a]

[a] Department of Animal Sciences, University of Wisconsin, Madison, WI 53706, USA
[b] Animal Production Research, Agricultural Research Centre – MTT, 31600 Jokioinen, Finland

(Received 14 May 1998; accepted 16 October 1998)

Abstract – Preferential treatment of cows in four herds of a multiple ovulation and embryo transfer scheme under selection was simulated. Prevalence and amount of preferential treatment depended on a function correlated with true breeding value. Three mixed effect linear models were compared in terms of their ability to handle preferential treatment: the classical Gaussian model, a model with multivariate t-distributed errors clustered by herd, and a model with independent t-distributed errors. In the models with t-distributed errors, both the scale parameters and the degrees of freedom were considered unknown. A Bayesian analysis was carried out for all three models via the Gibbs sampler, and posterior means were used to infer about genetic variance, herd-year effects, breeding values and realised response to selection. Performance over repeated sampling was assessed via Monte Carlo mean squared error. In the absence of preferential treatment, the three models had a similar performance. When preferential treatment was prevalent and strong, the univariate t-model was the best; hence, the Gaussian assumption for the errors was clearly inappropriate. It appears that some robust linear models can handle preferential treatment of animals better than the standard mixed effect linear model with Gaussian assumptions. © Inra/Elsevier, Paris

dairy cattle / preferential treatment / simulation / Bayesian statistics / Student-t distribution / Gibbs sampling

Résumé – **Atténuation des effets de traitement préférentiel dans un modèle linéaire mixte à distribution de Student (t). Étude de simulation.** On a simulé le traitement préférentiel de certaines vaches dans quatre troupeaux de sélection utilisant la transplantation embryonnaire. La fréquence et l'effet du traitement préférentiel

* Correspondence and reprints
E-mail: ismo.stranden@mtt.fi

ont dépendu d'une fonction corrélée à la valeur génétique vraie. On a comparé trois modèles linéaires mixtes pour leur aptitude à prendre en compte le traitement préférentiel : le modèle classique Gaussien, un modèle avec des erreurs t-multivariates groupées par troupeau et un modèle avec des erreurs t-distribuées indépendantes. Dans le modèle où les erreurs suivaient une distribution t, les paramètres d'échelle et les degrés de liberté ont été considérés inconnus. Une analyse bayésienne a été effectuée pour les trois modèles à partir de l'échantillonnage de Gibbs et les moyennes a posteriori ont été utilisées pour en inférer au sujet de la variance génétique, des effets troupeau-année, des valeurs génétiques et des réponses réalisées à la sélection. La performance des modèles a été évaluée au travers des erreurs quadratiques moyennes. En l'absence de traitement préférentiel, les trois modèles ont eu une performance similaire. Quand le traitement préférentiel a été fréquent et d'effet important, le modèle t-univariate a été le meilleur et le modèle Gaussien a été clairement inadapté. Il apparaît que des modèles linéaires robustes peuvent prendre en compte les traitements préférentiels mieux que les modèles linéaires mixtes Gaussiens classiques.

bovins laitiers / traitement préférentiel / simulation / statistique bayésienne / distribution de Student

1. INTRODUCTION

Preferential treatment is any management practice that is applied non-randomly to animals within a contemporary group [9]. For example, better housing and feeding, hormonal treatment, longer milking intervals on test day and feeding according to production are known to be applied selectively in dairy production. Preferential treatment occurs in dairy cattle, presumably to increase the economic value of a cow or the probability that it will be chosen as a bull-dam. Several studies (e.g. [17, 20]) have found that genetic evaluations for milk yield are inconsistent with expectations based on theory. This may be due to inadequate statistical assumptions or failure to account properly for selection or preferential treatment of cows.

Preferential treatment is often suspected when no apparent reasons exist for such discrepancies. Kuhn et al. [9] simulated effects of preferential treatment on 'animal model' genetic evaluations. Mean squared error of prediction of breeding values increased as the extent of preferential treatment increased. Kuhn and Freeman [10] found that when the dam of a sire was treated preferentially, more than 30 daughters with untreated records were needed to offset the bias in prediction of breeding value caused by the dam's information. Bias increased as the proportion and number of daughters receiving preferential treatment increased. Bias decreased when all daughters given preferential treatment were in the same herd; this is so because the 'herd-year' effect in the model captures part of the preferential treatment administered in a particular herd-year.

In order to account for preferential treatment, Harbers et al. [7] included an environmental correlation between related females in a genetic evaluation model for a MOET (multiple ovulation and embryo transfer) scheme. This improved accuracy of cow evaluations when preferential treatment was mild. Weigel et al. [29] simulated different strategies of preferential treatment and found that it was not possible to detect it by monitoring within-herd variance; obviously, this parameter does not provide information about the probability that a cow

within a herd is treated preferentially. Burnside and Meyer [3] simulated effects of bovine somatotropin (bST). Sire evaluations were least accurate when bST administration was targeted to the best producing cows.

In the context of prediction (e.g. [8]), a bias takes place when the expected values of the predictand and of the predictor differ. Evaluation of bias requires knowledge of the true model but, in practice, this is not available, so ad hoc assessments of bias have been suggested. Several studies [15, 16, 27, 28] found upward 'biases' of cow's pedigree indexes for protein or milk yield in Finnish Ayrshire. It is unclear if this discrepancy is due to chance, but preferential treatment of dams of cows may be a culprit. On the other hand, Powell and Norman [19] found that pedigree indexes understated the first estimated breeding values of daughters of proven sires mated to lower producing dams.

Little work has been undertaken on how to cope with preferential treatment in practice, at least from a statistical point of view. Kuhn and Freeman [11] studied power transformations of records but this was, at best, slightly effective in reducing bias due to preferential treatment. An alternative approach is to consider an error distribution with thicker tails than the normal, to allow for more variation. A commonly used one is the t-distribution, which is symmetric and leptokurtic. It has been advocated because of its simplicity [12], and because only one parameter (the degrees of freedom) is needed to describe robustness. A suitable robust distribution may be capable of attenuating the impact of outliers on data analysis. Many authors have employed statistical models with t-distributed residuals [4, 12, 13, 25, 31] in linear and non-linear regression models, with varying degrees of success. Use of the t-distribution in the context of mixed effects or hierarchical models is relatively recent [1, 2, 5, 6, 22–24, 26, 30].

Our objective was to assess frequentist properties of Bayesian point estimators obtained from mixed linear models where residuals were assumed to be either Gaussian or t-distributed. Milk production records obtained in herds in which some preferential treatment was practised were simulated. The analysis focused on mean squared error of estimation of genetic variance, herd-year effects, breeding values and genetic response to selection.

2. STRUCTURE OF THE SIMULATION

2.1. Conceptual population

Milk production records in a hypothetical 'adult' MOET nucleus scheme [18] were simulated. The scheme extended the simple hierarchical mating structure of Strandén et al. [21]. Our modification allowed bulls of the previous generation to mate current generation females. The nucleus consisted of 32 cows and eight bulls in every generation. In each generation, every nucleus cow produced (by multiple ovulation and embryo transfer to recipients) eight offspring, four females and four males. An animal could be selected only once into the nucleus as a parent and unselected animals were culled. The females were selected among those offspring to the nucleus that had completed a first lactation. Males were selected within those that had been born in the preceding generation. In practice, this would allow the bulls to have a progeny test outside the nucleus before selection. However, such progeny testing was not built in this simulation.

Thus, males within a full-sib family had the same estimated breeding value and three such males were randomly discarded. Each selected male was mated to four cows, chosen randomly from those that had been selected as replacements. Selection pressure in males and females was $\frac{8}{32} = \frac{1}{4}$ and $\frac{32}{128} = \frac{1}{4}$, respectively, per generation. With this scheme carried out for four generations, the data included 544 cows with records (32 in the base plus $32 \times 4 \times 4 = 512$ female progeny) and 32 sires with daughters in production, i.e. a total of 576 animals. A diagram of the simulated population is shown in *figure 1*.

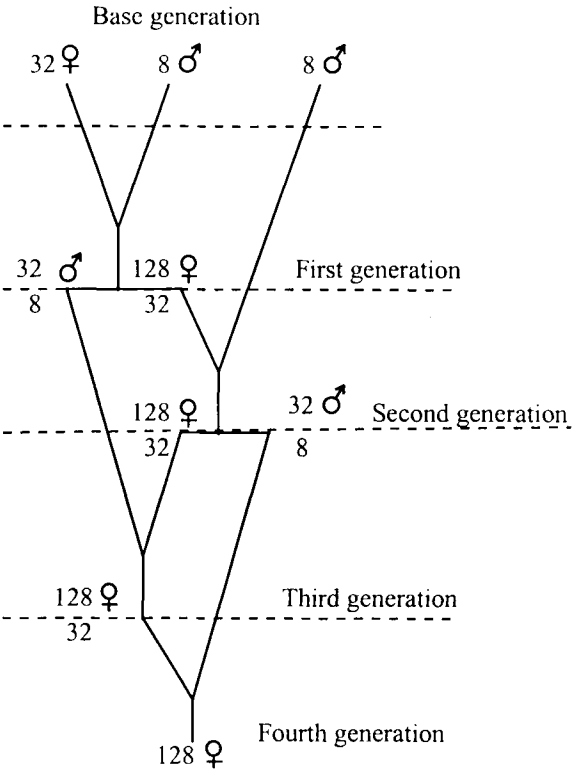

Figure 1. The hypothetical MOET nucleus breeding scheme.

Base generation cows were assigned to four herds in equal numbers, i.e. eight cows per herd. Female offspring of a cow remained in the same herd as her dam, whereas sires were used across herds. Breeding values of base animals were drawn at random from $N(0, 0.25)$ distribution. Records of the base animals were generated by adding a herd-year effect (independently, normally distributed) to a breeding value and to an independently drawn residual from $N(0, 0.75)$ distribution. Records in subsequent generations were simulated similarly, except that the breeding value of an individual was formed by averaging the breeding value of its parents and adding a $N\left(0, \frac{1}{2}\sigma_u^2(1 - \overline{F})\right)$

segregation residual, where σ_u^2 is the additive genetic variance and $\overline{F}$ is the average inbreeding coefficient of the parents. The selection criterion in the breeding scheme was BLUP of breeding value with the true variance components. The statistical model included the herd-year as a fixed effect and animal as a random effect (but ignored preferential treatment, as discussed later) using all genetic relationships available up to the time of selection.

2.2. Preferential treatment

In practice, preferential treatment takes place in the course of a selection programme so this is the way that the present simulation proceeded. None of the base population cows were treated preferentially, so there were 512 cows eligible to receive preferential treatment. A scheme in which the preferential treatment assigned depended on the 'perceived' breeding value of an animal (e.g. based on a genetic evaluation available before the animal produces the record) was adopted. The records were generated as

$$y_{ij} = h_i + u_j + e_{ij} + \Delta_{ij} \tag{1}$$

where y_{ij} is the record of animal j made in herd-year i, h_i is a herd-year effect, u_j is the breeding value of animal j, and e_{ij} is an independent residual. The preferential treatment Δ_{ij} was a stochastic effect taking the values:

$$\Delta_{ij} = \begin{cases} \Phi(w_j)(h_i - p_{\min}), & \text{if } w_j > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

where $\Phi(.)$ is the standard normal cumulative distribution function, $p_{\min}$ is a constant smaller than the herd-year effect h_i, and $w_j = \lambda + (u_j + v_j) \Big/ \sqrt{\sigma_u^2 + \sigma_v^2}$ is a 'value' function such that $u_i \sim N(0, \sigma_u^2)$, $v_j \sim N(0, \sigma_v^2)$, $\mathrm{Cov}(u_j, v_j) = 0$, so $w_j \sim N(\lambda, 1)$. In the preceding, σ_u^2 is the variance of breeding values and σ_v^2 is an 'uncertainty' variance. The ratio $\dfrac{\sigma_v^2}{\sigma_u^2}$ describes the uncertainty the herd manager has about the true breeding value of animal j. For example, if the breeder is very uncertain about the breeding value of the animal, this ratio of variances should be high. Three values of the uncertainty were considered, $\dfrac{\sigma_v^2}{\sigma_u^2} = \dfrac{1}{100}$, 1, 100. The correlation between w_j and the breeding value u_j is $\left(1 + \dfrac{\sigma_v^2}{\sigma_u^2}\right)^{-1/2}$, giving 0.995, 0.71 and 0.10 at values of the uncertainty equal to $\dfrac{1}{100}$, 1 and 100, respectively.

The preferential treatment scheme in equation (2) induces a correlation between related animals $\mathrm{Corr}(w_j, w_{j'}) = \dfrac{a_{jj'}\sigma_u^2 + \mathrm{Cov}(v_j, v_{j'})}{\sigma_u^2 + \sigma_v^2}$, where $a_{jj'}$ is the additive relationship between animals j and j'. If the v_j deviates are independent, then $\mathrm{Corr}(w_j, w_{j'}) = a_{jj'}\sigma_u^2/(\sigma_u^2 + \sigma_v^2)$. For example, if j and $j'$are full-sibs and $\sigma_v^2 = \sigma_u^2$, say, then $\mathrm{Corr}(w_j, w_{j'}) = \dfrac{1}{4}$. In general, the higher

the breeding value the higher the amount preferential treatment and the chance of receiving it.

The constant $p_{\min}$ in equation (2) controls the range of production associated with preferential treatment. It was set equal to $-5\sigma_h$ where σ_h is the standard deviation of herd-year effects. These were drawn from a normal distribution with mean zero and variance σ_h^2, and two different values of the herd-year variance were considered: $\sigma_h^2 = \sigma_u^2$ and $\sigma_h^2 = 3\sigma_u^2$. The constant λ controls the proportion of cows to be preferentially treated. Normal distribution theory can be used to find a value of λ such that a desired proportion of cows receives preferential treatment. The proportion of preferentially treated cows increases with λ, because $\Pr(w_j > 0)$ increases concomitantly. Three different prevalences of preferential treatment were considered: 1 out of 10, 1 out of 32, and 1 out of 64 cows. These correspond to λ values of -1.2816, -1.8627 and -2.1539, respectively.

It was intended to keep the proportion of preferentially treated animals roughly constant from generation to generation. To do so, it must be noted that selection is expected to increase mean breeding value and to reduce genetic variance over time. In order to account for these effects, the formula for w was changed to:

$$w_j = \lambda + (u_j - \overline{u} + v_j) \Big/ \sqrt{\sigma_v^2 + S_u^2} \tag{3}$$

where $\overline{u}$ is the mean breeding value of animals available for preferential treatment in the generation to which animal j belongs, and S_u^2 is the additive genetic variance for individuals born in that generation.

The probability distribution of the amount of preferential treatment (Δ_{ij}) depends on the values of σ_h and λ as shown in the Appendix. The average amount of preferential treatment actually applied was assessed via a simulation of 1 000 replicates of the MOET scheme. Mean increase (mean of Δ) in production due to preferential treatment under varying prevalence of preferential treatment and amount of herd-year variance is in *table I*. As intended, production increased with prevalence of preferential treatment, and with σ_h^2. Average value of preferential treatment was not affected by level of uncertainty $\frac{\sigma_v^2}{\sigma_u^2}$. This is not shown in *table I*, but it was expected because the distribution of Δ_{ij} does not depend on this ratio.

Table I. Average increase in simulated lactation production due to preferential treatment as a function of herd-year variance (σ_h^2) and of prevalence of preferential treatment (values in parenthesis are Monte Carlo standard errors from 1 000 replicates of the MOET scheme, σ_u^2 = additive genetic variance).

	Prevalence of preferential treatment		
	1/10	1/32	1/64
$\sigma_h^2 = \sigma_u^2$	1.67 (0.21)	1.58 (0.34)	1.37 (0.61)
$\sigma_h^2 = 3\sigma_u^2$	2.89 (0.37)	2.75 (0.57)	2.38 (1.05)

2.3. Statistical models and computations

Three linear statistical models were compared, both with and without preferential treatment incorporated in the simulation. The objective was to assess the relative ability of these models to handle perturbations caused by unknown preferential treatment. In all three models, the linear structure for the records included an unknown herd-year effect (treated as fixed computationally), the unknown breeding value of the cow and a residual, distributed according to an appropriate error distribution, as noted below. In the three models, a multivariate normal distribution $N(\mathbf{0}, \mathbf{A}\sigma_u^2)$, where $\mathbf{A}$ is a 576×576 relationship matrix, was used for the genetic effects, so there was no difference in this respect. The three models, differing only in the error distribution were the following.

1) G: a purely Gaussian model with errors Niid$(0, \sigma_e^2)$.

2) t-1: errors were independently and identically distributed as univariate-t, $t_1(0, \sigma_e^2, v_e)$. Here, the variance of the distribution is $\sigma_e^2 v_e/(v_e - 2)$, where σ_e^2 is a scale parameter and v_e are the unknown degrees of freedom.

3) t-H: within herd i ($i = 1, 2, 3, 4$), the error vector $\mathbf{e}_i$ had the multivariate-t distribution $t_{n_i}(\mathbf{0}, \mathbf{I}_{n_i}\sigma_e^2, v_e)$ where n_i is the number of records in herd i. Here, $\text{Var}(\mathbf{e}_i) = \mathbf{I}_{n_i}\sigma_e^2 v_e/(v_e - 2)$. Although the errors are uncorrelated, they are not independent, this being a property of the multivariate t-distribution. Error vectors in different herds were mutually independent, however, but with the same σ_e^2 and v_e parameters. We refer to this model as a 'herd-clustered' one.

The G model is the usual one; model t-1 discounts outliers y_{ij} on a 'case' by 'case' basis, and model t-H discounts outlying vectors $\mathbf{y}_i$ for the entire herd i. Because the 'value function' w_j used to generate preferential treatment does not depend on the herd, there is no apparent reason why model t-H should outperform model t-1. It should be noted that as $v_e \to \infty$, the two t-distributions tend towards the Gaussian one.

A Bayesian structure was adopted for inference. Prior distributions were the same for all three models. Herd effects were assigned a uniform prior and, as noted, a multivariate normal process was used as a prior distribution for the breeding values. The dispersion components σ_u^2 and σ_e^2 were assigned independent scaled inverted chi-square distributions with four degrees of freedom and mean equal to the true variance component, i.e. 0.75 for the residual variance and 0.25 for the genetic variance. In the t-models, the prior for σ_e^2 is for the scale of the distribution and not for the residual variance, which is $\sigma_e^2 v_e/(v_e - 2)$ as noted before. In the two models involving the t-distribution, the residual degrees of freedom parameter v_e was considered unknown. Degrees of freedom values allowed in the herd-clustered t-model were 4, 10, 100 or 1 000, all equally likely, a priori. In the univariate t-distribution model, the space of v_e was 4, 6, 8, 10, 12 or 14, all receiving equal prior probability. These values were chosen arbitrarily. It is possible to use a continuous prior for v_e [23] but the discrete distribution employed here facilitated implementation. A Gibbs sampler was used to carry out the Bayesian computations employing the full conditional distributions described in Strandén [22]. Tests made in several simulations with varying starting values indicated that a burn-in period of 7 000 iterates with 70 000 Gibbs iterates thereafter (all samples kept) was enough to obtain sufficiently precise estimates of posterior means of the parameters.

About 60 min of CPU time were required to perform 70 000 iterations, for any of the models, in an HP 9 000(3) computer.

2.4. Frequentist comparison

Each replicate of the simulation consisted of a data set generated as per the scheme in *figure 1* under the appropriate assumptions of preferential treatment. A Bayesian analysis of the data set according to each of the three models was carried out in each replicate. Mean squared errors of posterior mean estimates were computed, over replicates, for: a) genetic variance, b) herd-year effects, and c) breeding values. Mean squared errors were also computed for three classes of breeding values: sires, cows who had been preferentially treated and cows without preferential treatment. d) An additional end-point of interest was mean squared error of estimated response to selection, assessed by predicting breeding values using posterior means from the three models contrasted. 'True' response was the mean difference in true breeding value (due to selection using BLUP) between animals born in the last generation and those born in the first generation. Differences in mean squared errors between models should reflect the relative accuracy of estimation of genetic trend.

A 'pilot run' [14] was conducted to assess the number of replicates needed to attain enough precision for a parameter of interest. The approximate number of replications required to achieve an absolute precision r for the confidence interval given a pilot run of n replicates was found using:

$$N(r,n) = \min\left\{ i \geq n | t_{i-1,1-\alpha/2}\sqrt{\frac{S_n^2}{i}} \leq r \right\} \tag{4}$$

where $t_{i-1,1-\alpha/2}$ is the value of a t-distribution with $i-1$ degrees of freedom at the $100(1-\alpha)$ percentile ('confidence'). Our pilot study consisted of carrying $n = 20$ replicates for each of the three models. The number of replications required to achieve 0.05 precision with 95 % confidence for the genetic variance was less than 60 for most cases. Hence, it was decided that all cases would be replicated 60 times. Absolute precision was recalculated after 60 replicates, and a further 40 replicates were made for the schemes involving 1/10 prevalence of preferential treatment. One scheme $\left(\frac{\sigma_h^2}{\sigma_u^2} = 3, \frac{\sigma_v^2}{\sigma_u^2} = \frac{1}{100}\right)$ required an additional 40 replicates to achieve the required precision. *Table II* indicates the schemes and number of replicates performed.

Because of its heavy computing requirements, the analysis was performed using a network of machines administered by Professor Miron Livny of the Department of Computer Science, University of Wisconsin at Madison. This cluster was accessed using the Condor system, which allows running jobs simultaneously at many computers while the data and program reside in one computer. Each replicate of each model was a process to be executed in this network of computers. There were between 10 and 15 computers available at any time, giving at least a 10-fold increase in computing power compared to using only the HP9000(3).

Table II. Number of replicates performed under different prevalences of preferential treatment (PT), level of uncertainty $\left(\frac{\sigma_v^2}{\sigma_u^2}\right)$ and ratio of herd-year variance to additive genetic variance $\left(\frac{\sigma_h^2}{\sigma_u^2}\right)$.

PT	$\frac{\sigma_v^2}{\sigma_u^2}$	$\frac{\sigma_h^2}{\sigma_u^2}$	Number of replicates
1/64	$\frac{1}{100}$, 1, 100	1, 3	60
1/32	$\frac{1}{100}$, 1, 100	1, 3	60
1/10	$\frac{1}{100}$, 1, 100	1	100
1/10	1,100	3	100
1/10	$\frac{1}{100}$	3	140

3. RESULTS AND DISCUSSION

3.1. Absence of preferential treatment

The objective here was to examine possible losses in efficiency due to using the two t-distribution models when there is no preferential treatment and the Gaussian assumption holds throughout. Averages and mean squared errors of estimates of additive genetic variance are given in *table III*. The posterior means of σ_u^2 for each of the three models were practically unbiased, in light of the Monte Carlo variation. However, the mean squared error was larger for the two t-models than for the Gaussian one. Hence, if the Gaussian assumption holds, posterior means of additive genetic variance for the t-models are less accurate than those from the G-model. The increase in mean squared error over the Gaussian model was about 5–6 % for the t-H model, and 7–18 % for the t-1 model.

Tables IV and *V* give the posterior distributions of the degrees of freedom for the two t-models in the absence of preferential treatment. The analysis carried out with the herd-clustered t-model clearly favoured a model with Gaussian errors, as indicated by a posterior probability of about 90 % for the degrees of freedom being larger than 10. Also, the univariate t-model assigned the highest posterior probability, about 40 %, to the largest value of the degrees of freedom ($v_e = 14$) considered. The posterior distributions were not sharp, this being a function of the low informational content the data have about v_e. However, both analyses favoured the larger values of v_e or, equivalently, the Gaussian assumption for the errors. For example, in the herd-clustered t-model, the posterior odds ratio of $v_e = 1\,000$ relative to $v_e = 4$ was 17.7 and 29:7 for $\frac{\sigma_h^2}{\sigma_u^2} = 1$ and $\frac{\sigma_h^2}{\sigma_u^2} = 3$, respectively. In the univariate t-model, the odds ratio

Table III. Average (over replicates) and mean squared error of posterior means of additive genetic variance, by model, in the absence of preferential treatment (Monte Carlo standard errors in parenthesis).

σ_h^2/σ_u^2	Model G	t-H	t-1	True value
1) Average				
1	0.22 (0.011)	0.23 (0.011)	0.25 (0.012)	0.25
3	0.23 (0.011)	0.23 (0.011)	0.26 (0.012)	
2) Mean squared error				
1	0.0073	0.0077	0.0078	
3	0.0071	0.0075	0.0084	

σ_u^2 = additive genetic variance, σ_h^2 = variance between herd-years, G = Gaussian model, t-H = herd-clustered t-model, t-1 = univariate t-model.

Table IV. Posterior distribution of the degrees of freedom for the herd-clustered t-model in the absence of preferential treatment.

	Degrees of freedom			
σ_h^2/σ_u^2	4	10	100	1 000
1	0.028	0.089	0.389	0.495
3	0.018	0.064	0.384	0.534

σ_h^2 = variance between herd-years, σ_u^2 = additive genetic variance.

Table V. Posterior distribution of the degrees of freedom for the univariate t-model in the absence of preferential treatment.

	Degrees of freedom					
σ_h^2/σ_u^2	4	6	8	10	12	14
1	0.001	0.030	0.101	0.194	0.290	0.384
3	0.001	0.022	0.088	0.187	0.297	0.404

σ_h^2 = variance between herd-years, σ_h^2 = additive genetic variance.

of $v_e = 14$ relative to $v_e = 4$ was 384 and 404 for the two values of the ratio between herd and additive genetic variances.

Mean squared errors of estimates of location parameters were similar in all models (*table VI*), although slightly smaller for the G-model. As expected, mean squared errors were larger for breeding values of cows (smallest amount of information) than for sires. When herd-year variance was large, relative to the additive genetic variance, mean squared error of estimation of breeding values increased. When estimating realised response to selection, the mean squared errors were 0.031 (G model), 0.030 (t-H model) and 0.029 (t-1 model).

In summary, in the absence of preferential treatment and with the Gaussian assumption holding throughout, the t-models were less accurate for estimation of σ_u^2, but were as competitive as the Gaussian model for estimation of breeding values and of genetic trend.

Table VI. Mean squared errors of posterior mean estimates of herd-year effects (HY), all breeding values (BV), breeding values of sires (BV_{Sire}) and breeding values of cows (BV_{Cow}) in the absence of preferential treatment.

		Model		
	σ_h^2/σ_u^2	G	t-H	t-1
HY	1	0.068	0.069	0.069
	3	0.075	0.075	0.075
BV	1	0.164	0.164	0.165
	3	0.175	0.175	0.176
BV_{Sire}	1	0.127	0.127	0.127
	3	0.152	0.153	0.153
BV_{Cow}	1	0.166	0.167	0.167
	3	0.176	0.177	0.177

σ_u^2 = additive genetic variance, σ_h^2 = variance between herd-years, G = Gaussian model, t-H = herd-clustered t-model, t-1 = univariate t-model.

3.2. Preferentially treated data

3.2.1. Additive genetic variance

Mean squared error of estimates of additive genetic variance are in *figure 2*. Differences between models were clearest when preferential treatment was more prevalent (1/10) and when the herd-year variance was high (this affects the distribution of Δ). Also, differences between models were largest when uncertainty about true breeding values was low, so the value function is a high correlate of breeding value. There was little difference between the G and the t-H models, but the univariate t-model had the best performance when prevalence of preferential treatment was medium (1 out of 32 cows) or high (1 out of 10 cows). The univariate t-model was robust to variation in the uncertainty parameter; this was not the case for the G and the t-H models, whose performance was hampered under severe forms of preferential treatment.

3.2.2. Posterior distribution of the degrees of freedom

Posterior probabilities of the degrees of freedom under the herd-clustered t-model were often higher for the larger values of v_e, thus supporting the Gaussian model, especially when preferential treatment was uncommon, or uncertainty was high. Only under medium (1/32) or high (1/10) prevalence of preferential treatment and a high herd-year variance the largest values of the degrees of freedom did not have the highest posterior probability. However,

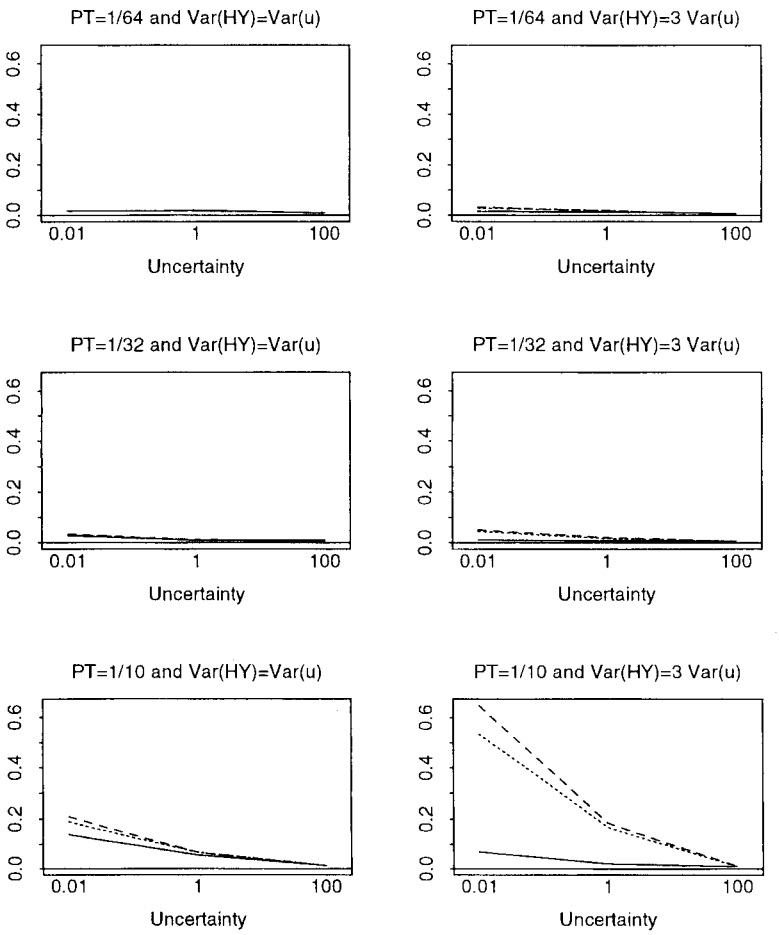

Figure 2. Mean squared error of estimates of additive genetic variance, by model, under different prevalence of preferential treatment (PT), uncertainty $\left(\frac{\sigma_v^2}{\sigma_u^2}\right)$ and amount of herd-year (HY) variance. Dashed line = Gaussian, dotted line = herd-clustered t-model, and solid line = univariate t-model.

this depended on the level of uncertainty and on the amount of herd-year variance. For example, when prevalence of preferential treatment was 1/10 and with $\sigma_h^2 = 3\sigma_u^2$, low values of the degrees of freedom had higher posterior probabilities when uncertainty was low; however, as uncertainty increased, the posterior distribution tended to favour larger values of the degrees of freedom.

Posterior probabilities of v_e for the univariate t-model are given in *figure 3*. Here, posterior distributions tended to be flat. Higher probabilities were assigned to the largest values of the degrees of freedom only when preferential treatment was rare and the herd-year variance low. As in the herd-clustered t-model, high uncertainty often resulted in higher probabilities assigned to the highest degrees of freedom values, as one would expect. However, other degrees

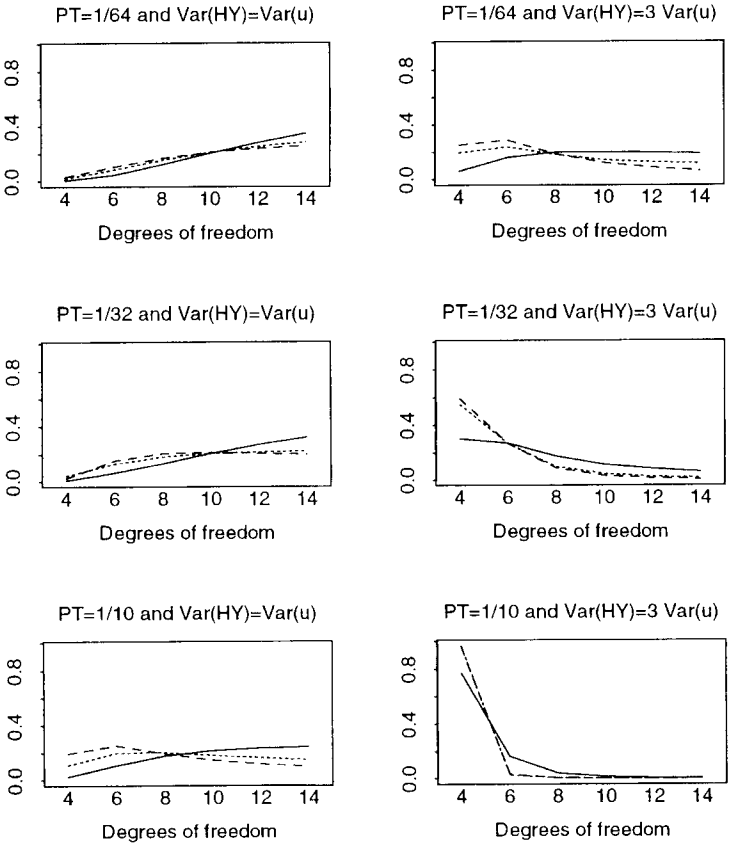

Figure 3. Posterior probabilities of different values of the residual degrees of freedom in univariate t-model under different uncertainties $\left(\frac{\sigma_v^2}{\sigma_u^2}\right)$, prevalence of preferential treatment (PT) and amount of herd-year (HY) variance. Dashed, dotted and solid lines are used to indicate uncertainties of $\frac{1}{100}$, 1 and 100, respectively.

of freedom values also received relatively high probabilities. When preferential treatment was prevalent (1/10) and the herd-year variance was large, the posterior distribution was sharp, with a modal value of $v_e = 4$ at all levels of uncertainty. This points away from a Gaussian distribution of the residuals. With a small data set such as the one in this simulated MOET scheme, one should not expect the posterior distribution of the degrees of freedom parameter to be highly peaked. Nevertheless, the univariate t-model recognised the non-Gaussian situation even when prevalence was rare (1/64), provided that the variance between herds was relatively large. This is because the expected value of the preferential treatment, $\mathrm{E}(\Delta_{ij})$, increased with σ_h^2, as illustrated in *table I.*

3.2.3. *Estimates of herd-year effects, breeding values and genetic response to selection*

Average of mean squared error of estimates of herd-year effects was similar for the three models except when preferential treatment was prevalent or herd-year variance was high, but it was always smallest for the univariate t-model (*figure 4*). When preferential treatment was common (1/10), the univariate t-model clearly had the smallest mean squared error at each level of uncertainty and value of σ_h^2.

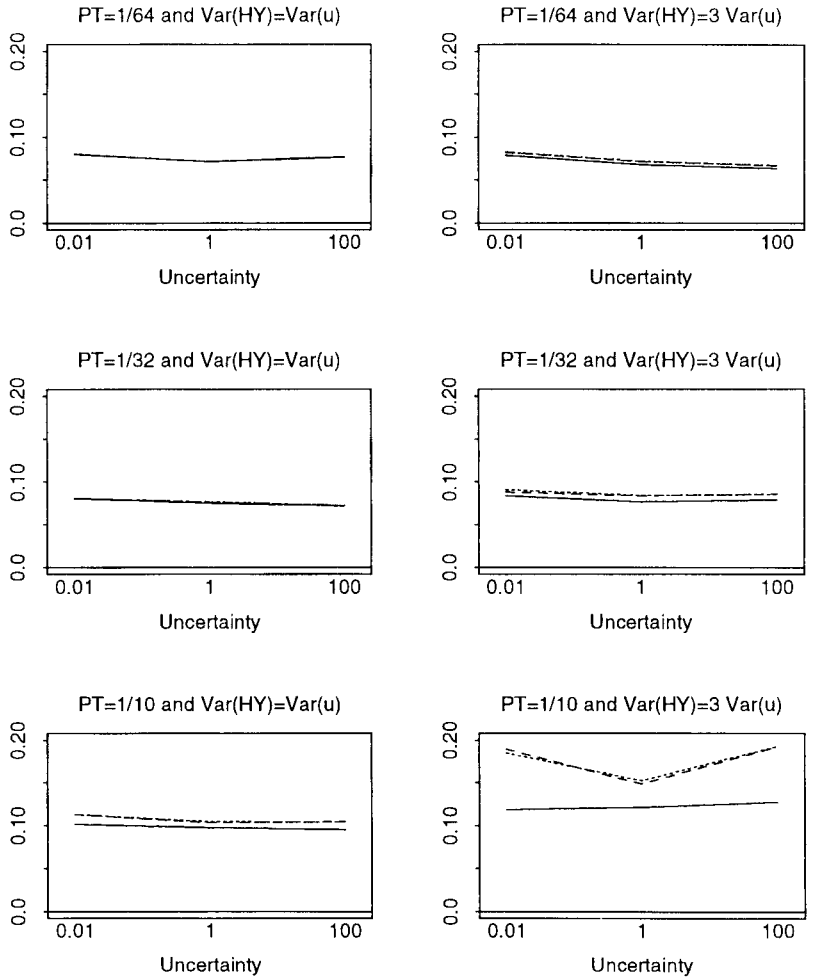

Figure 4. Mean squared error of estimates of herd-year effects, by model, under different prevalence of preferential treatment (PT), uncertainty about breeding values and amounts of herd-year (HY) variance. Dashed line = Gaussian, dotted line = herd-clustered t-model and solid line = univariate t-model.

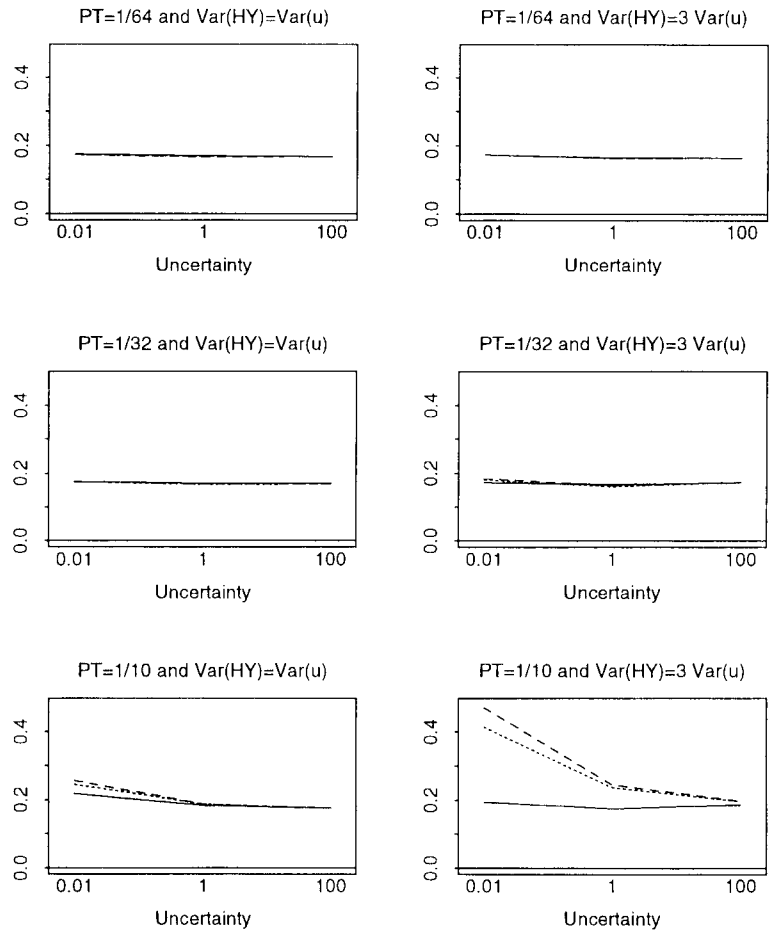

Figure 5. Mean squared error of estimates of breeding values, by model, under different prevalence of preferential treatment (PT), uncertainty about breeding values and amounts of herd-year (HY) variance. Dashed line = Gaussian, dotted line = herd-clustered t-model and solid line = univariate t-model.

The average of mean squared error of estimates of all breeding values is shown in *figure 5*. This criterion was about the same with all models except when preferential treatment was common and the herd-year variance high. Here, when uncertainty was high, there were no differences between the models, but at low levels of uncertainty, the univariate t-model was markedly superior. The picture for mean squared errors of estimates of sire and cow breeding values and for genetic response was similar to that for of all breeding values, so the figures are not presented. In all cases, differences between models were clear, favouring the univariate t-model when preferential treatment was more prevalent (1/10). The same was true for preferentially treated cows (*figure 6*), but mean squared errors were larger than for breeding values of cows that were not treated preferentially. The univariate t-model had a similar or slightly worse

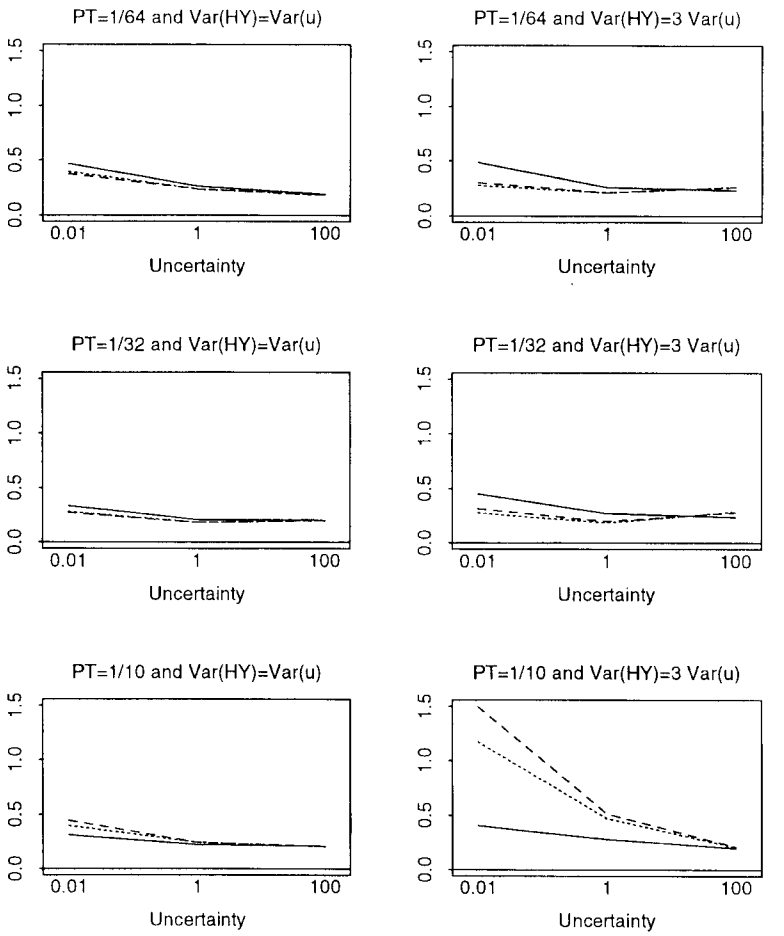

Figure 6. Mean squared error of estimates of preferentially treated cows' breeding values, by model, under different prevalence of preferential treatment (PT), uncertainty about breeding values and amounts of herd-year (HY) variance. Dashed line = Gaussian, dotted line = herd-clustered t-model and solid line = univariate t-model.

performance than the Gaussian or herd-clustered models when preferential treatment was rare or mildly prevalent, but it was superior when such treatment was common. In particular, at the lowest level of uncertainty and at the highest herd-year variance, the univariate t-model gave predictions of breeding value of preferentially treated cows that had a mean squared error of about a third of that observed with the Gaussian model. In this situation, the herd-clustered model improved estimates somewhat relative to the Gaussian model.

4. CONCLUSIONS

In the absence of preferential treatment, the t-models were as good as the Gaussian model for estimating breeding values and response to selection. When preferential treatment was mildly prevalent (1/32) the models performed similarly. However, when preferential treatment was common (1/10) and especially when the herd-year variance was large relative to the additive genetic variance, the univariate t-model was clearly the best, at least in terms of mean squared error. Under preferential treatment, the posterior distribution of the degrees of freedom in the univariate t-model pointed away from the correctness of the Gaussian assumption. The univariate t-model was quite robust to variation in the simulation parameters, but it is unknown whether this robustness holds across different forms of preferential treatment.

This simulation could not differentiate clearly between the Gaussian and the herd-clustered t-models, although the latter was always slightly better under preferential treatment. A reason for the lack of difference between these two models may be the low number of herds in the simulation. With a few clusters (herds) the statistical information about the degrees of freedom is low, so the posterior distribution of this parameter cannot be estimated accurately.

In conclusion, it appears that the univariate t-model can attenuate adverse effects of preferential treatment as applied here. It leads to better inferences about breeding values and genetic trends than those obtained with the Gaussian model, especially when preferential treatment is prevalent, at least under the conditions of the study. If, on the other hand, preferential treatment is nonexistent, or the assumption of a Gaussian distribution of the residuals seems to be true, there is little loss in efficiency from using a robust model, such as the univariate t. It is encouraging that a symmetric error distribution, such as Student t, improved upon the Gaussian one under a single-tailed form of preferential treatment as in equation (2). This suggests that a robust asymmetric distribution may do even better, but perhaps at the expense of conceptual and computational simplicity.

ACKNOWLEDGEMENT

We wish to thank W.G. Hill, University of Edinburgh, for some useful comments.

REFERENCES

[1] Albert J.H., Chib S., Bayesian analysis of binary and polychotomous response data, J. Am. Stat. Assoc. 88 (1993) 669–679.

[2] Besag J., Green P., Higdon D., Mengersen K., Bayesian computation and stochastic systems, Stat. Sci. 10 (1995) 3–66.

[3] Burnside E.B., Meyer K., Potential impact of bovine somatotropin on dairy sire evaluation, J. Dairy Sci. 71 (1988) 2210–2219.

[4] Geweke J., Bayesian treatment of the independent Student-t linear model, J. Appl. Econometrics 8 (1993) S19–S40.

[5] Gianola D., Strandén I., Foulley J.L., Modelos lineales con distribuciones t: potencial en genetica quantitativa, Actas, 5ta Conferencia Espanola de Biometria, Valencia, Spain, 1995, pp. 3–4.

[6] Gianola D., Sorensen D., A mixed effects threshold model with a t distribution, 47th Annual Meeting of the European Association for Animal Production, Lillehammer, Norway, 1996, 15 p.

[7] Harbers A.G.F., Lohuis M.M., Dekkers J.C.M., Correction for preferential treatment of MOET families by including an environmental correlation in genetic evaluation, in: Proceedings of the 5th World Congress on Genetics Applied to Livestock Production, Guelph, Canada, 1994, vol. 17, pp. 11–14.

[8] Henderson C.R., Applications of Linear Models in Animal Breeding, University of Guelph, Guelph, Ontario, Canada, 1984.

[9] Kuhn M.T., Boettcher P.J., Freeman A.E., Potential biases in predicted transmitting abilities of females from preferential treatment, J. Dairy Sci. 77 (1994) 2428–2437.

[10] Kuhn M.T., Freeman A.E., Biases in predicted transmitting abilities of sires when daughters receive preferential treatment, J. Dairy Sci. 78 (1995) 2067–2072.

[11] Kuhn M.T., Freeman A.E., Power transformations for reducing bias in genetic evaluation caused by preferential treatment, J. Dairy Sci. (Abstr.) (1996) suppl. 1, 143.

[12] Lange K.L., Little R.J.A., Taylor J.M.G., Robust statistical modeling using the t distribution, J. Am. Stat. Assoc. 84 (1989) 881–896.

[13] Lange K., Sinsheimer J.S., Normal/Independent distributions and their applications in robust regression, J. Comp. Graph. Stat. 2 (1993) 175–198.

[14] Law A.M., Kelton W.D., Simulation Modeling and Analysis, McGraw-Hill, New York, 1982.

[15] Lidauer M., Mäntysaari E.A., Detection of bias in animal model pedigree indices of heifers, Agric. Food Sci. Finland 5 (1996) 387–397.

[16] Mäntysaari E.A., Sillanpää M., Bias in pedigree indices of dairy bulls: Should the management group effects be fixed and should we use smaller heritability?, 44 th Annual Meeting of the European Association for Animal Production, Aarhus, Denmark, 1993, Abstracts I, pp. 236–237.

[17] Murphy P.A., Everett R.W., Van Vleck L.D., Comparison of first lactations and all lactations of dams to predict sons' milk evaluation, J. Dairy Sci. 65 (1982) 1999–2005.

[18] Nicholas F.W., Smith C., Increased rates of genetic change in dairy cattle by embryo transfer and splitting, Anim. Prod. 36 (1983) 341–353.

[19] Powell R.L., Norman H.D., Accuracy of cow indexes according to repeatability, evaluation, herd yield, and registry status, J. Dairy Sci. 71 (1988) 2232–2240.

[20] Rothschild M.F., Douglass L.W., Powell R.L., Prediction of son's modified contemporary comparison from pedigree information, J. Dairy Sci. 64 (1981) 331–341.

[21] Strandén I., Mäki–Tanila A., Mäntysaari E.A., Genetic progress and rate of inbreeding in a closed adult MOET nucleus under different mating strategies and heritabilities, J. Anim. Breed. Genetics 108 (1991) 401–411.

[22] Strandén I., Robust mixed effects linear models with t distributions and application to dairy cattle breeding, Ph.D. thesis, University of Wisconsin, Madison, 1996.

[23] Strandén I., Gianola D., Gaussian versus Student-t mixed effects linear models for milk yield in Ayrshire cattle, 48th Annual Meeting of the European Association for Animal Production, Vienna, 1997, Abstracts 1, pp. 262–263.

[24] Strandén I, Gianola D., Inferences about variance components in the univariate mixed linear t model using Laplacian-t approximations, in: Proceedings of the 6th World Congress on Genetics Applied to Livestock Production, Armidale, Australia, 1998, vol. 25, pp. 537–540.

[25] Sutradhar B.C., Ali M.M., Estimation of the parameters of a regression model with a multivariate t error variable, Comm. Stat. Theory Meth. 15 (1986) 429–450.

[26] Tempelman R.J., Firat M.Z., Beyond the linear mixed model: perceived versus real benefits, Proceedings of the 6th World Congress on Genetics Applied to Livestock Production, Armidale, Australia, 1998, vol. 25, pp. 605–612.

[27] Uimari P., Mäntysaari E.A., Repeatability and bias of estimated breeding values for dairy bulls and bull dams calculated from animal model evaluations, Anim. Prod. 57 (1993) 175–182.

[28] Uimari P., Mäntysaari E.A., Relationship between bull dam herd characteristics and bias in estimated breeding values of bull, Agric. Sci. Finland 4 (1995) 463–472.

[29] Weigel D.J., Pearson R.E., Hoeschele I., Impact of different strategies and amounts of preferential treatment on various methods of bull-dam selection, J. Dairy Sci. 77 (1994) 3163–3173.

[30] West M., Outlier models and prior distributions in Bayesian linear regression, J. Roy. Statist. Soc. B 46 (1984) 431–439.

[31] Zellner A., Bayesian and non-Bayesian analysis of the regression model with multivariate Student-t error terms, J. Am. Stat. Assoc. 71 (1976) 400–405.

APPENDIX: Distribution of the preferential treatment variable

When w_j is positive and very large, Δ_{ij} tends to $h_i - p_{\min}$, so in this case equation (1) becomes:

$$y_{ij} = 2h_i - p_{\min} + u_j + e_j \tag{A.1}$$

When w_j is negative, $\Delta_{ij} = 0$, as indicated in (1) so $y_{ij} = h_i + u_j + e_j$. Hence, given h_i, the range in production records due to preferential treatment is expected to be $h_i - p_{\min} = h_i + 5\sigma_h = (z_i + 5)\sigma_h$ where $z_i \sim N(0,1)$. Unconditionally, the expected range is then $5\sigma_h$. For Δ_{ij} defined in equation (2), the average preferential treatment applied, conditionally on h_i would be

$$\begin{aligned} \mathrm{E}(\Delta_{ij}|h_i) &= \int_{-\infty}^{0} 0\phi_\lambda(w)\mathrm{d}w + \int_{0}^{\infty} \Phi(w)(h_i - p_{\min})\phi_\lambda(w)\mathrm{d}w \\ &= (h_i - p_{\min}) \int_{0}^{\infty} \Phi(w)\phi_\lambda(w)\mathrm{d}w \end{aligned} \tag{A.2}$$

where $\phi_\lambda(.)$ is normal density with mean λ and variance 1. For $\lambda = 0$, $\phi_\lambda(.)$ is the standard normal density $\phi(.)$. Because $\int_{-\infty}^{z} \Phi(w)\phi(w)\mathrm{d}w = \frac{1}{2}\Phi^2(z)$ and $\Phi(0) = \frac{1}{2}$, it follows that

$$\begin{aligned} \mathrm{E}(\Delta_{ij}|h_i) &= \frac{1}{2}(h_i - p_{\min})[1 - \Phi^2(0)] \\ &= \frac{3}{8}(h_i - p_{\min}) \end{aligned} \tag{A.3}$$

so $\mathrm{E}(\Delta_{ij}) = \mathrm{E}(\mathrm{E}(\Delta_{ij}|h_i)) = -\frac{3}{8}p_{\min}$. Likewise, $\mathrm{E}(\Delta_{ij}^2|h_i) = \frac{7}{24}(h_i - p_{\min})^2$ for $\lambda = 0$. Thus,

$$\mathrm{Var}(\Delta_{ij}) = \frac{29}{192}p_{\min}^2 + \frac{56}{192}\sigma_h^2 \tag{A.4}$$

With $p_{\min} = -5\sigma_h$, we have $\mathrm{E}(\Delta_{ij}) = 1.875\sigma_h$ and $\mathrm{Var}(\Delta_{ij}) \approx 4.068\sigma_h^2$. Then, C.V.$(\Delta_{ij}) \approx 108$ % when 50 % of the cows receive preferential treatment.