Random Regression Models for Production Traits in Canadian Holsteins

J. Jamrozik[1], L.R. Schaeffer, and J.C.M. Dekkers

Centre for Genetic Improvement of Livestock, Department of Animal and Poultry Science
University of Guelph, Guelph, ON, N1G 2W1, Canada

Introduction

Test day (TD) yields for production traits are the components for calculation of standard 305-d lactation yields in dairy cattle which are subsequently used for genetic evaluation. Test day models (TDM) have been proposed for modelling TD yields directly. TDM can account for factors that are specific to each test day, such as management groups within a herd on test day, day of the year, number of days in milk (DIM), pregnancy status, medical treatment, and number of times milked on test day. Many of these factors can change for a cow from one test day to the next, and would be difficult to model for 305-d yields.

Ptak and Schaeffer (1993) used a repeated records animal model that included covariates to describe the general shape of lactation curve within fixed age-season of calving subclasses for genetic evaluation of TD yields in first lactation. Resulting estimated breeding values (EBV) measured genetic differences in the height of curves between animals. The multiple trait version of this model has been applied for genetic evaluation of somatic cell scores from the first three lactations in Canada (Reents et al., 1995). Schaeffer and Dekkers (1994) presented an extension of the TDM by allowing the shape of lactation curve to differ for each cow at the genetic level. This was accomplished by the inclusion of random regression coefficients for each animal. The lactation curve for an individual cow was modelled based on two sets of regressions on DIM. Fixed regressions for all cows belonging to the same subclass describe the average shape for that group of cows, and the random regressions for a cow describe the deviations from the fixed regressions. This model for TD yields is called a random regression test day model (RRM).

Variance and covariance components for RRM were estimated by Jamrozik and Schaeffer (1996) for first lactation milk, fat, and protein yields of Canadian Holsteins. RRM was then used for genetic evaluation of 1.1 million Canadian Holstein cattle based on 5.1 million first lactation TD records (Jamrozik et al., 1996). Linear and quadratic functions on DIM and log of DIM were used to describe both fixed and random regressions in these models. Each animal evaluated received several EBV for various parts of lactation. Three measures of persistency of lactation were compared, and it was concluded that animals could be selected for both high yields and high persistency of their daughters (Jamrozik et al., 1996).

The shape of lactation curve can be modelled by different functions, which means that different random regression models are possible. Functions which can be used for the RRM should be linear with respect to the coefficients to be estimated, should not involve too many parameters, should provide a good fit for both average lactation curves (fixed regressions) and random deviates from average lactation curves (random regressions). Different functions can be used to describe fixed versus random regressions.

The objective of this study was to compare three RRM for variance component estimation and genetic evaluation of milk and protein yields in Canadian Holsteins. Two functions were used to model fixed and random

[1] on leave from Department of Genetics and Animal Breeding, Agricultural Academy, al. Mickiewicza 24/28, 30-059 Krakow, Poland

regressions in RRM. They were: linear and quadratic functions on DIM and log of DIM (Ali and Schaeffer, 1987) and a sum of linear and exponential functions on DIM, which was used by Wilmink (1987) for fitting lactation curves.

Models

The general form of the single trait, single lactation RRM for dairy production traits can be written as

$$y_{ijkl} = MG_i + f_k(DIM_{jl}) + r_j(DIM_{jl}) + p_j + e_{ijkl}$$

where

- y_{ijkl} is daily yield on test day l for cow j in management group i and fixed lactation curve subclass k
- DIM_{jl} is the day of lactation for cow j on test l
- MG_i is the fixed effect of management group i
- p_j is the random environment effect common to all test days of cow j
- e_{ijkl} is the effect of random environment specific to y_{ijkl}
- f_k is a function of DIM with unknown fixed coefficients that describe the average shape of lactation curves for cows in group k
- r_j is a function of DIM with unknown random genetic coefficients that describe deviates from f_k which are specific for the genetic effects of cow j.

To assure linearity of the model, both f_k and r_j should be linear with respect to unknown parameters. In this case $f_k(t) = \Sigma \beta_{km} w_m(t)$, $r_j(t) = \Sigma \gamma_{jm} u_m(t)$, where w_m and u_m are known functions, β_{km} are fixed regression coefficients within subclass k, and γ are random genetic coefficients specific to cow j, with covariance structure G.

The model can then be written in matrix form as

$$\mathbf{y} = \mathbf{X}_{MG}\mathbf{b}_{MG} + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{W}\mathbf{p} + \mathbf{e}, \text{ with}$$

$$var\begin{bmatrix} \gamma \\ p \\ e \end{bmatrix} = \begin{bmatrix} G \otimes A & 0 & 0 \\ 0 & I\sigma^2_p & 0 \\ 0 & 0 & R \end{bmatrix}$$

where

γ, p and e are vectors of genetic, permanent environment, and residual effects, respectively,
A is the numerator relationship matrix, $G=var[\gamma_{j1}, \gamma_{j2}, \ldots, \gamma_{jn}]'$,
R is the covariance matrix of residual terms,
σ^2_p is the variance of permanent environment.

Two functions were used in this study to describe fixed and random regressions, namely

$$R(t) = a_0 + a_1*(t/305) + a_2*(t/305)^2 + a_3*\log(305/t) + a_4*[\log(305/t)]^2$$

(Ali and Schaeffer, 1987) and

$$W(t) = a_0 + a_1*t + a_2*\exp(-0.05*t)$$

(Wilmink, 1987)

Three RRM were set up with R(t) and W(t) as fixed and(or) random regression. The models were:

R/R model: $f_k = R$, $r_j = R$,
R/W model: $f_k = R$, $r_j = W$,
W/W model: $f_k = W$, $r_j = W$.

Management groups were defined as herd-test date (HTD) subclasses. Groups for fixed regression were created by region-age at calving-season at calving subclasses. R was assumed to be diagonal with elements determined by DIM to account for differences in residual variance by stage of lactation (29 classes of DIM: 5-20, 21-30, 31-40,....,281-290, 290-305).

Material and methods

The initial data included TD records on Holstein cows with daily milk and protein yield (kg) from 4 milk recording organization in Canada (Ontario, Quebec, Prairies and British Columbia), for cows calving from 1988 through 1995. Data were restricted to first lactation, TD records had DIM between 5 and 305 days, and age at first calving had to be between 18 and 48 months. Edits included milk yield in the range 1.5-90 kg, and protein % in the range 1.5-10%.

(co) Variance components

Due to computational limitations, a subset of the data that included 181 randomly chosen herds from the two largest regions (Ontario and Quebec), was used for estimation of (co)variance components for the three RRM. Data included 50,412 TD records on 6,763 cows made in 6,757 HTD subclasses. Four age classes (18-24, 25-29, 30-34 and 35-48 months) and two seasons of calving (September-February, March-August) formed sixteen region-age of calving-season of calving subclasses. The pedigree file contained 13,912 animals (12,544 cows and 1,368 bulls).

Gibbs sampling was used to generate marginal posterior distributions of co(variance) components for all three models. Variance and covariance components for random effects in RRM were then estimated as simple means of 25,000 generated samples. Details of distributional assumptions for the RRM and the computing algorithm for Gibbs sampling are in Jamrozik and Schaeffer (1996).

To compare the different RRM, daily genetic variances were estimated for each DIM from 5 to 305 as

$$g_i = \mathbf{z}_i' \, \mathbf{G} \, \mathbf{z}_i$$

were

$\mathbf{z}_i$ is a vector of covariates for the i-th DIM for either R(t) or W(t). Error variances (29 estimates for each trait/model combination) and variance of permanent environment could be compared directly between models.

Genetic evaluation

A total of 5,101,635 first lactation TD records on 709,357 cows were used for genetic evaluations based on R/R, R/W and W/W models. There were 13,322 sires and 567,493 HTD subclasses represented in the data. Including ancestors, the total number of animals evaluated was 1,096,272. Region-age-season subclasses (32 in total) were formed in a similar way as for variance components estimation.

The models used were the same as for co(variance) components estimation but modified by including phantom parent group effects. Six groups were formed on the basis of sire and dams pathways and year of birth for females. Mixed model equations were set up similar to Quaas (1988), where group plus animal additive genetic effects were solved directly. Iteration on data with the Gauss-Seidel method was used to solve the mixed model equations. Iterations were performed until either the sum of squares of changes between rounds divided by the sum of squares of the latest solution attained a value of less than 1×10^{-8} or the number of iterations exceeded 350. A group of bulls with at least 50 daughters with data was chosen for comparison of evaluations from different models.

Solutions for random regression coefficients were used to generate EBV for various (part)lactation yields in the following way: let $\mathbf{z}_i$ be the vector of random regression covariates for the i-th DIM and **a** represent a vector of sums of solutions for group effects plus additive genetic effects for an animal. EBV of the animal for yield on the i-th day of lactation can be calculated as $g_i = \mathbf{z}_i'\mathbf{a}$. EBV for 305-d lactation can be obtained by summing EBV for each DIM from 1 to 305. EBV for part lactation yields can be obtained by varying the interval of the summation. EBV for 305-d yield (T) and three part lactation yields (T_1 - days 1-100, T_2 - days 101-200, T_3 - days 201-305) were calculated for milk and protein

yields. EBV for persistency (P) was defined as the additional genetic yield (gained or lost) from day 60 to 280, relative to an animal with average persistency and the same yield on day 60, and was computed as $P = (g_{280} - g_{60})*110$.

The accuracy and precision of RRM were assessed by comparing means and variances of residuals of the models which were estimated for each test day.

Results and discussion

(co) Variance components

Estimates for genetic variance-covariance matrix **G** were not comparable between R/R and models involving the Wilmink function but comparisons can be made for daily genetic variances which can be calculated as values of quadratic forms of **G**. Plots of genetic variances for DIM from 5 to 305 from R/R, W/W and R/W models are shown in Figures 1 and 3 for milk and protein, respectively. Figures 2 and 4 present estimates of error variance from RRM. Estimates of permanent environment variances are shown in Table 1.

There was generally good agreement between models in estimates of co(variance) components for both analysed traits. Both models with the Wilmink function gave practically identical daily genetic variance across the whole lactation. Differences were largest between models R/R and W/W models on DIM 5 to 15 for milk yield (Figure 1). Daily genetic variances showed a certain degree of curvilinearity along the DIM scale. The largest values of genetic variance were observed at the beginning of lactation. For all models the genetic variance was virtually stabilized in the middle part of lactation, with a subsequent slight increase at the end of lactation.

Estimates of error variances were also very similar between models (Figures 2, 4), although some differences were found at the beginning and at the end of lactation. Error variance for the first interval (DIM 5-20) was much larger for model R/R than for either W/R or W/W. An opposite trend was observed for the last interval (DIM 290-305). This may suggest that models with Wilmink curve as random regression better describe deviates from standard lactation curve at the beginning of lactation, whereas the opposite holds for the last part of lactation.

Estimates of permanent environment variance for protein yield were the same for all models (Table 1). Some (not larger than 0.5 kg^2) differences were observed for milk yield.

All models gave very similar estimates of daily phenotypic variances (sum off all components discussed above) for both traits, indicating equal ability of R(t) and W(t) functions in describing the variability of TD yields. Comparison of variance components estimates from the R/R model with estimates obtained by Jamrozik and Schaeffer (1996) for the same traits and the same model but with a different (disjoint) data set showed a very high degree of similarity. This indicates the invariance of estimates for the R/R model to the sampling of data.

Genetic evaluation

Differences in EBV from the R/R, W/W and R/W models are summarized in Table 2, for a group of 1,730 bulls with at least 50 daughters with data. All values are expressed relatively to the standard deviations of EBV from the R/R model for the analysed group of sires. Standard deviations of EBV for T, T1, T2, T3 and P were 578, 171, 213, 227, 180 and 15.8, 4.5, 5.8, 6.6, 5.4 (kg^2) for milk and protein, respectively.

R/R and W/W models gave on average the most similar estimates for total lactation milk yield, followed by R/W - W/W and R/R - R/W combinations. The R/W model overestimated T on average by 50 kg of milk yield in comparison with both R/R and R/W models. For protein yield, R/R - R/W differences were slightly larger than those from other comparisons. It is difficult to outline the general pattern of differences between models in part lactation EBV. Different parts of lactation showed different behaviour with respect to different traits.

Relative differences for T1, T2 and T3 were larger than those for T. The largest values of standard deviations of differences can be observed for the second and the third 100d of lactation. Mean relative differences in EBV for persistency were significantly larger than those for (part)lactation EBV.

Correlations of T, T1, T2, T3 and P between models for bulls with 50 or more daughters were all greater than 0.99, indicating almost identical ranking of sires by R/R, W/W and R/W models. Correlations of T with official Canadian January 1996 evaluations from a multiple lactation repeatability animal model were equal to 0.94 for both analysed traits.

Accuracy of models can be further examined by analyzing estimates of its residuals. Figures 5 and 7 show plots of mean residuals from the R/R, W/W and R/W models computed on a daily basis for milk and protein, respectively. Absolute values of residuals were smaller than 0.5 kg for milk yield and smaller than 0.03 kg for protein yield. Differences between models were small and largest at the beginning and at the end of lactation. Models R/R and R/W generated the same values for mean daily residuals of protein yield. Model R/R seemed to give on average the smallest bias for both traits, except days 5-20 for milk yield. The model with the Wilmink function as both fixed and random regressions produced significantly larger daily residual for the last part of lactation (DIM>250) for both traits, and for DIM 20 - 70 for protein yield. This may indicate that function R is more accurate than the Wilmink function in modelling average daily yields with RRM.

Precision of models can be assessed by the variance of estimated residuals. Daily residual variances from the R/R, W/W and R/W models for milk and protein yields are shown in Figures 6 and 8, respectively. Values of variances were in the range of daily error variance components, discussed earlier in this paper. The shape of graphs was similar for all models/traits combinations. The largest values of variances were observed around peak lactation for milk yield (day 50) and at the end of the lactation. Model R/R gave significantly smaller residual variances for DIM greater than 250 and DIM in the range from 20 to 70 than models R/W and W/W. Models with Wilmink functions were more precise in the beginning part of lactation (DIM from 5 to 20). No differences in precision were observed for W/W and R/W models, except DIM 280-305 for protein yield.

Daily residual variances from RRM can be compared with variances of residuals estimated by regular TD model (without random regressions) from the study of Ptak and Schaeffer (1994). Their estimates for milk yield were in the range of 3.5 - 9.5 kg^2, which indicates the superiority of RRM over TD, mainly at the beginning and at the end of lactation.

Model R/R required 3 min. 20 sec. of CPU time on HP 9000/735 workstation with 130 MB of memory for one round of iteration with 6,800,800 equations. It took 314 and 320 rounds of iteration to achieve the assumed convergence criterion with model R/R for milk and protein, respectively. A summary of computing requirements of other RRM (relative to R/R model) is shown in the Table 3. Since the number of equations was smaller for models with the Wilmink function (5 versus 3 equations per animal), models R/W and W/W required less memory and less CPU time per iteration. Model R/W showed poorer convergence behaviour for both traits, reaching the value of $1x10^{-6}$ in round 350.

It could be possible (but was not done in this study) to examine the behaviour of the W/R model (W for fixed, R for random regression). Based on the current comparison it could be speculated, however, that the accuracy, precision and required computing resources of W/R will be very close to those of R/R model.

Conclusions

The first application of RRM for dairy production traits used the R function for both fixed and random regressions (Jamrozik et al., 1996). The present study was undertaken to compare the properties of the R/R model with two alternative models involving linear functions of smaller number of parameters.

The general behaviour of RRM was similar across all analysed combinations of fixed/random regression, although the performance of individual models differed slightly between traits. Ranking of bulls was the same for all three models. Model R/R was on average more accurate and precise than either model W/W or R/W, especially for protein yield. Models with the Wilmink function (W/W or W/R) could be used in the situations when computing resources are limited. It could be possible that other combinations of functions (for example the Wilmink function with a coefficient for the exponential component different from -0.05) might give a better fit to TD records (smaller residuals and residuals squared). This study used milk and protein yields of Canadian Holsteins as data for testing RRM. Different populations (traits) may require different functions for the optimal modelling of random and fixed regressions.

Acknowledgments

Financial support from the Ontario Ministry of Agriculture , Food and Rural Affairs, the Cattle Breeding Research Council of Canada, and the Natural Sciences and Engineering Research Council is gratefully acknowledged.

References

Ali, T.E. and Schaeffer, L.R. 1987. Accounting for covariances among test day milk yields in dairy cows. Can. J. Anim. Sci. 67, 637.

Jamrozik, J. and Schaeffer, L.R. 1996. Estimates of genetic parameters for a test day model with random regressions for production of first lactation Holsteins. J. Dairy. Sci. (submitted)

Jamrozik, J., Schaeffer, L.R. and Dekkers, J.C.M. 1996. Genetic evaluation of dairy cattle using test day yields and a random regression model. J. Dairy Sci. (submitted)

Ptak. E. and Schaeffer, L.R. 1993. Use of test day yields for genetic evaluation of dairy cattle. Livest. Prod. Sci. 34, 23.

Quaas, R.L. 1988. Additive genetic model with groups and relationships. J. Dairy Sci. 71, 1338.

Reents, R., Dekkers, J.C.M. and Schaeffer, L.R. 1995. Genetic evaluation for somatic cell score with a test day model for multiple lactation. J. Dairy Sci. 77, 2671.

Schaeffer, L.R. and Dekkers, J.C.M. 1994. Random regressions in animal models for test-day production in dairy cattle. Proc. 5th World Congr. Genet. Appl. Livest. Prod., Guelph, vol. XVIII, 443.

Wilmink, J.B.M. 1987. Adjustment of test-day milk, fat and protein yields for age, season and stage of lactation. Livest. Prod. Sci. 16, 335.

Table 1. Estimates of permanent environment variance for milk and protein yield (kg) from R/R, W/W and R/W models

	R/R	W/W	R/W
Milk yield	5.8	5.7	5.3
Protein yield	0.005	0.005	0.005

Table 2. Means and standard deviations of differences between EBV from the R/R, W/W and R/W models, expressed in SD of EBV from the R/R model (x100), for (part) lactation yields (T), and persistency (P) of milk and protein yield (kg) (for 1,730 bulls with more than 50 daughters)

		Milk		Protein	
		Mean	SD	Mean	SD
R/R - W/W	T	-0.8	4.8	-2.7	9.2
	T1	-3.3	9.2	+3.6	9.3
	T2	-0.7	10.8	-1.0	13.3
	T3	+1.1	10.2	-8.0	17.9
	P	+3.4	15.4	-16.2	32.0
R/R - R/W	T	-8.5	3.2	-4.2	4.7
	T1	+2.9	7.2	-0.9	8.2
	T2	-30.9	10.6	-13.4	12.2
	T3	+5.2	9.2	+0.2	8.6
	P	77.3	13.8	+34.1	13.6
R/W - W/W	T	+7.7	2.6	+1.6	6.6
	T1	-6.2	4.4	+4.5	6.0
	T2	+30.2	2.6	+12.4	4.5
	T3	-4.1	3.3	-10.2	14.1
	P	-73.9	6.1	-50.3	28.1

Table 3. Computing resources for models W/W and R/W relative to model R/R (in %)

		W/W	R/W
No of equations		68	68
CM (iteration program)		81	83
CPU time/iteration		59	59
No of iterations to reach convergence	Milk	96	>115
	Protein	101	>119

Figure 1: Estimates of genetic variances of daily milk yield (kg) from models R/R, W/W and R/W

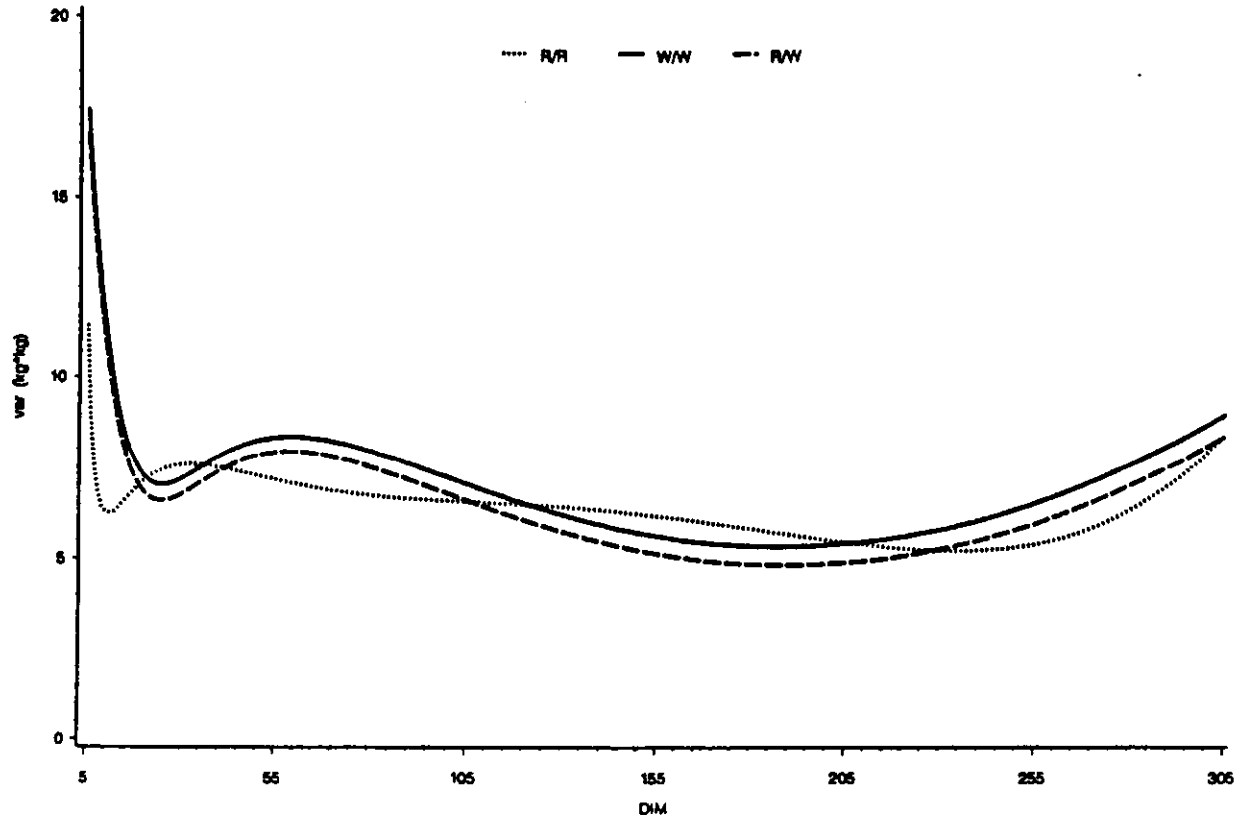

Figure 2: Estimates of error variances of daily milk yield (kg) from models R/R, W/W and R/W

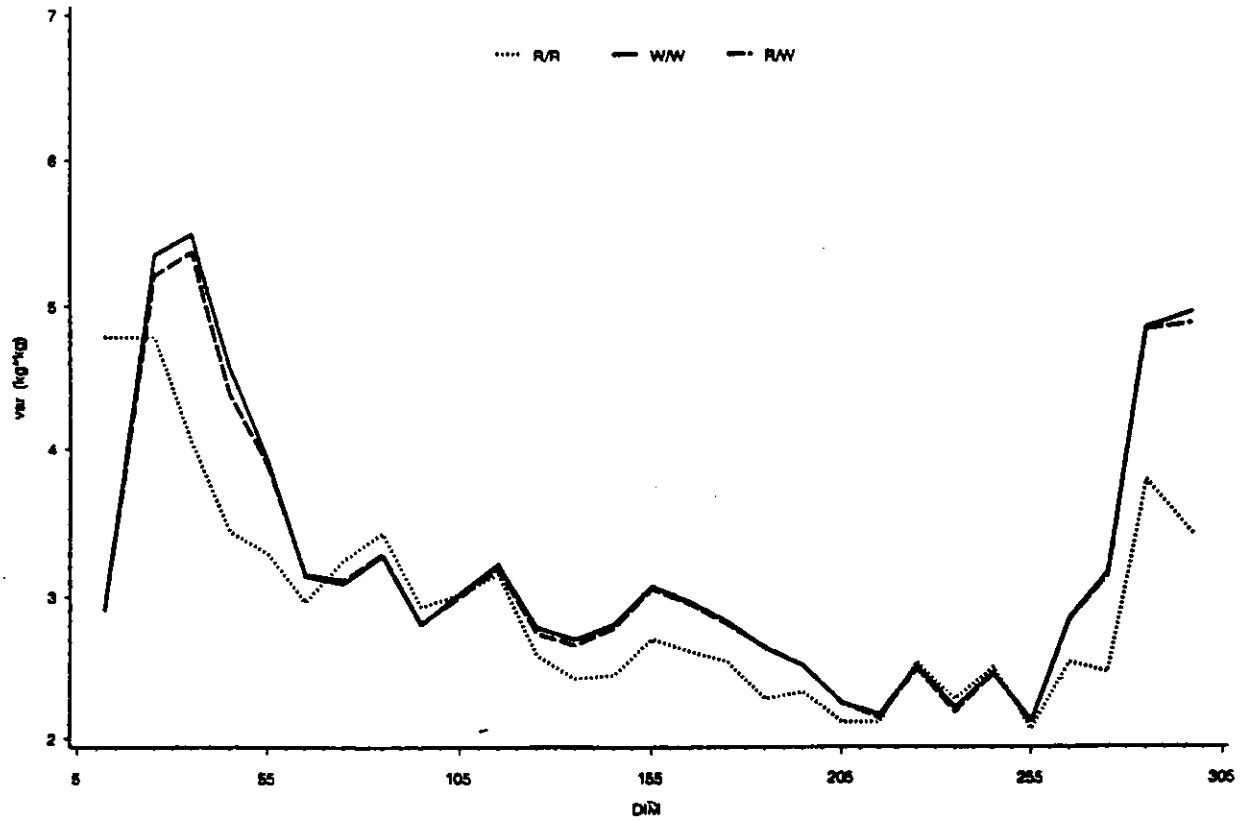

Figure 3: Estimates of genetic variances of daily protein yield (kg) from models R/R, W/W and R/W

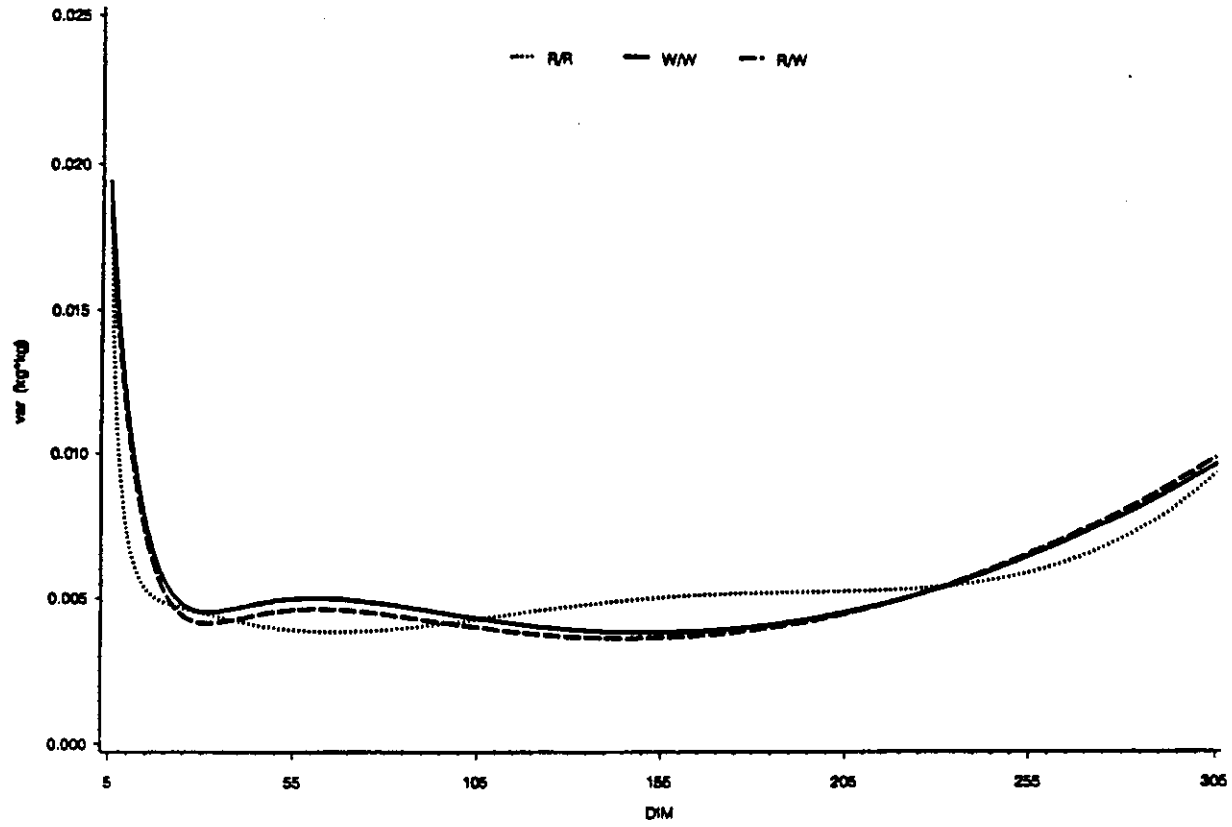

Figure 4: Estimates of error variances of daily protein yield (kg) form models R/R, W/W and R/W

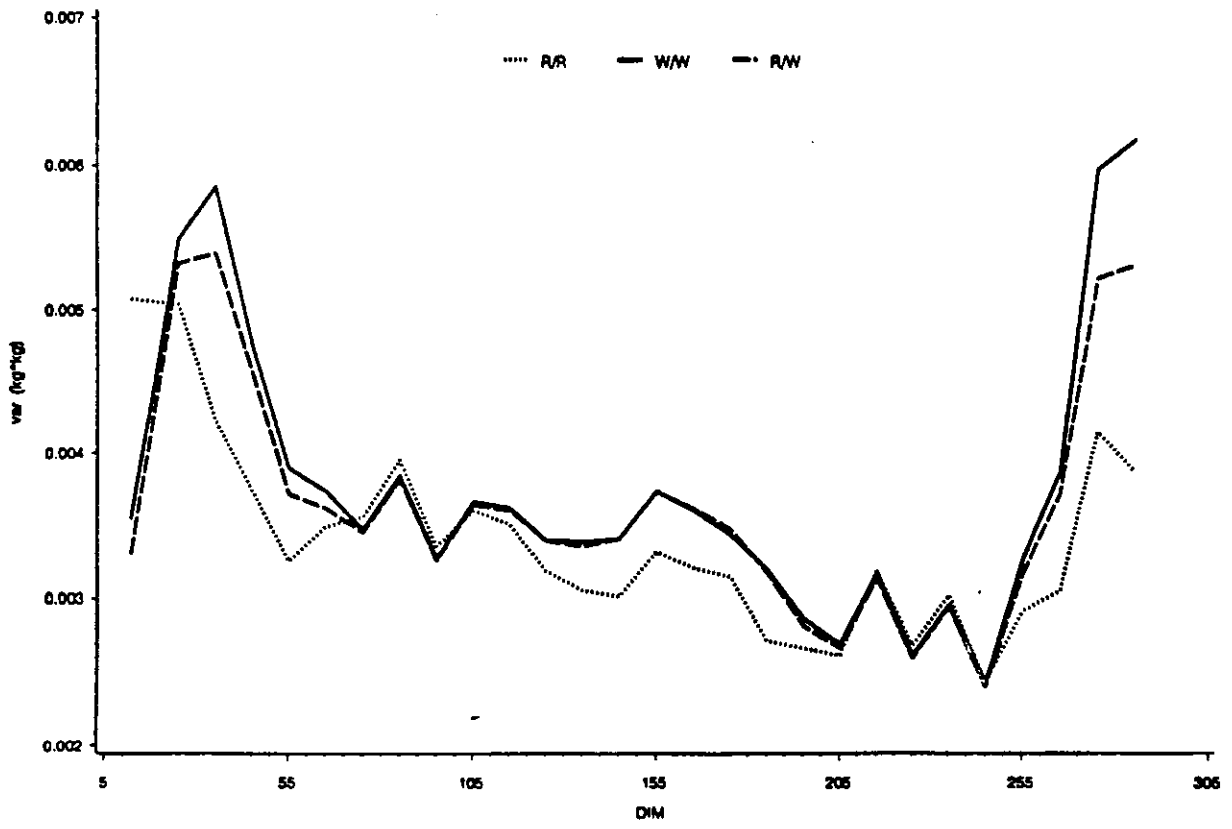

Figure 5: Mean residuals of daily milk yield (kg) from models R/R, W/W and R/W

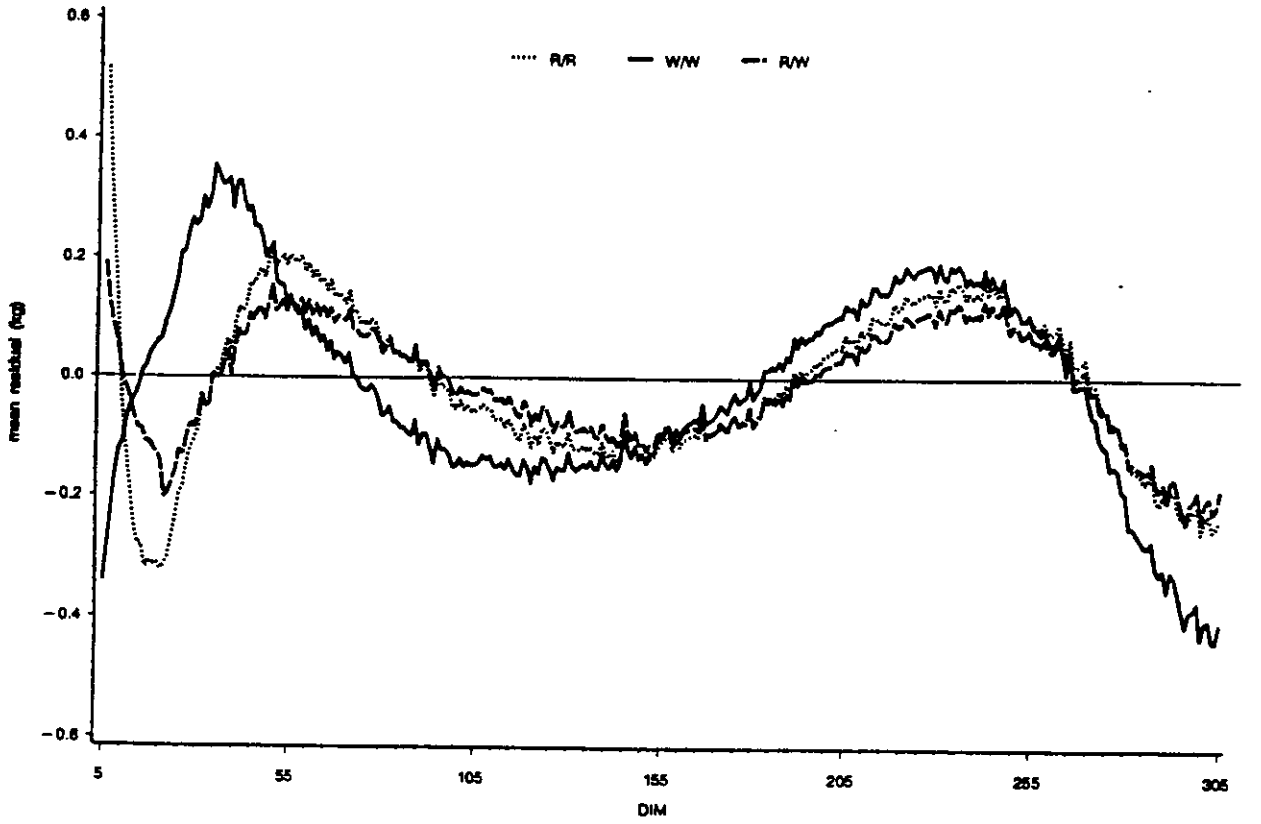

Figure 6: Residual variances of daily milk yield (kg) form models R/R, W/W and R/W

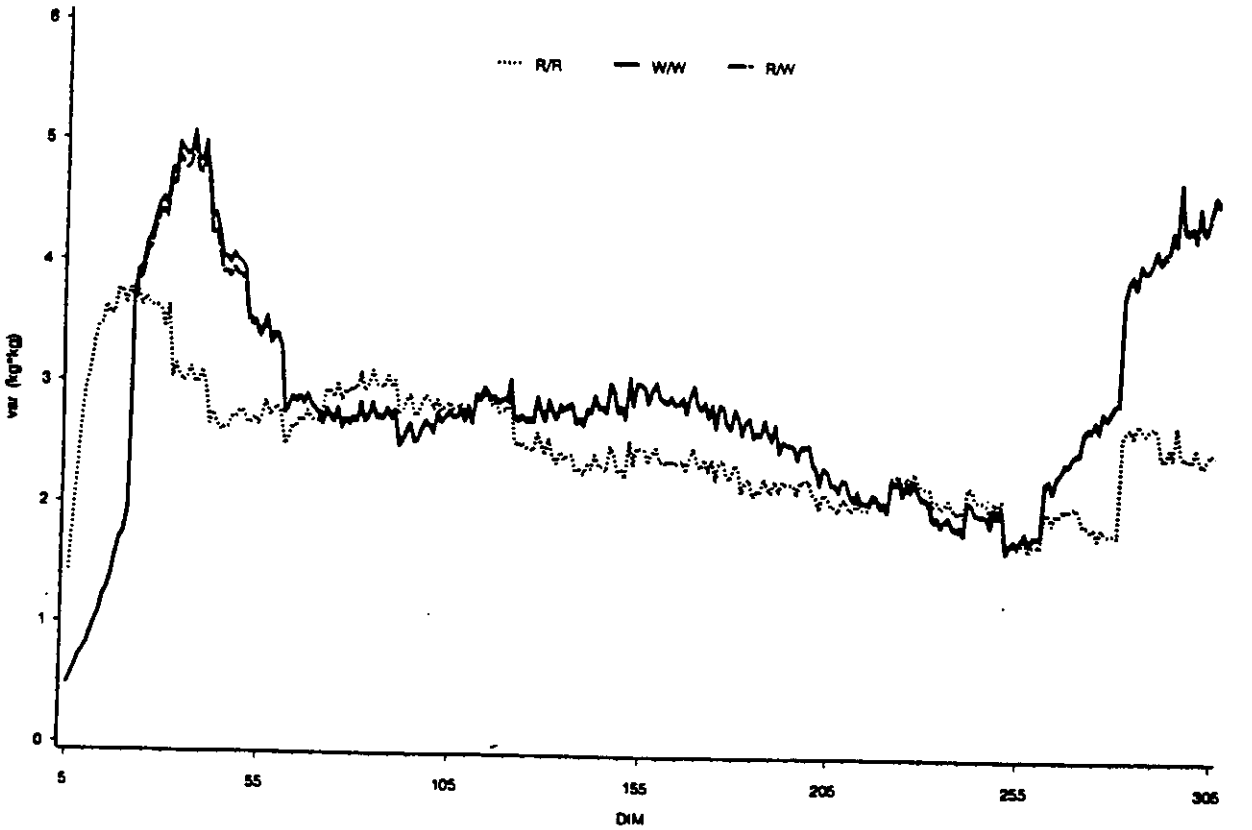

Figure 7: Mean residuals of daily protein yield (kg) from models R/R, W/W and R/W

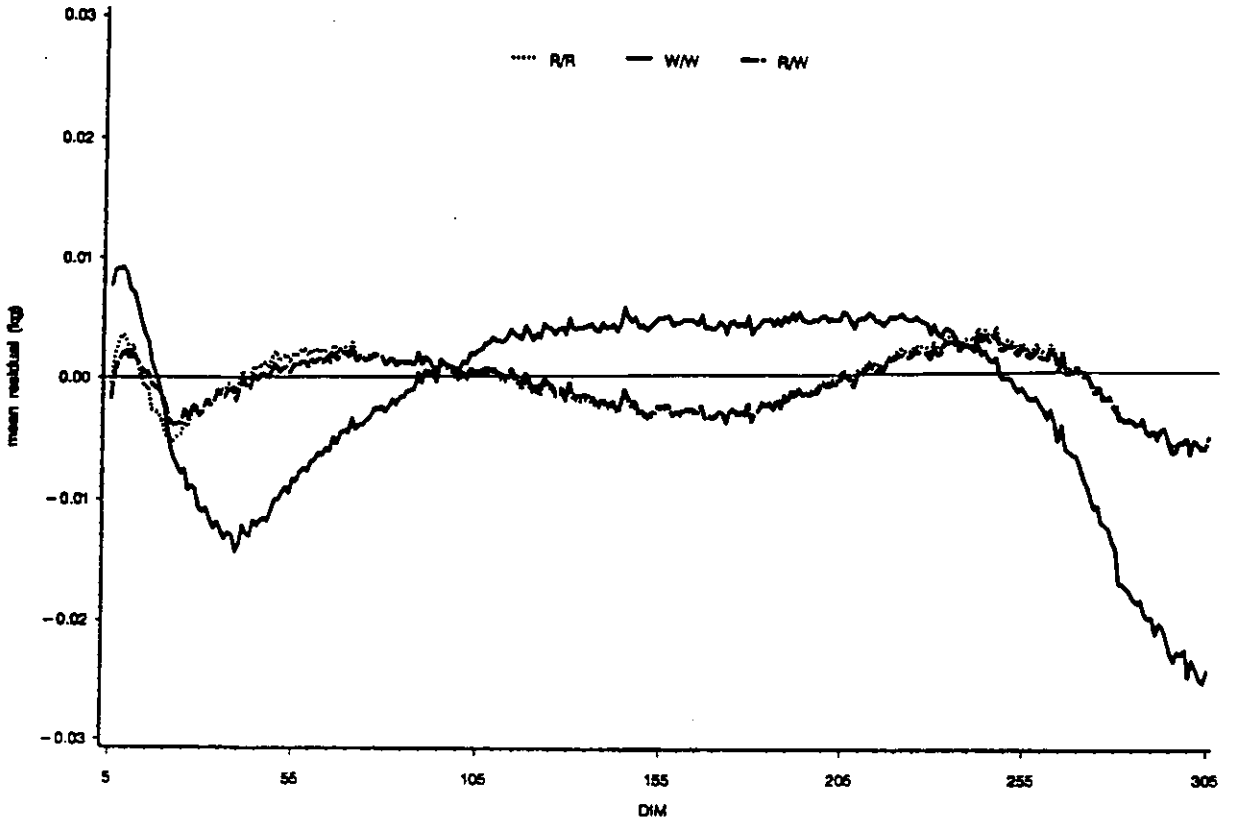

Figure 8: Residual variances of daily protein yield (kg) from models R/R, W/W and R/W