A Link Function Approach to Model Heterogeneity of Residual Variances Over Time in Lactation Curve Analyses

F. Jaffrezic,* I.M.S. White,* R. Thompson,† and W. G. Hill*

*Institute of Cell Animal and Population Biology, University of Edinburgh, West Mains Rd., Edinburgh EH9 3JT, UK
†Rothamsted Experimental Station, IACR, Harpenden, Herts AL5 2JQ, UK
Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS, UK

ABSTRACT

Several studies with test-day models for the lactation curve show heterogeneity of residual variance over time. The most common approach is to divide the lactation length into subclasses, assuming homogeneity within these classes and heterogeneity between them. The main drawbacks of this approach are that it can lead to many parameters being estimated and that classes have to be arbitrarily defined, whereas the residual variance changes continuously over time. A methodology that overcomes these drawbacks is proposed here. A structural model on the residual variance is assumed in which the covariates are parametric functions of time. In this model, only a few parameters need to be estimated, and the residual variance is then a continuous function of time. The analysis of a sample data set illustrates this methodology.

(**Key words:** lactation curves, random regression model, heterogeneity of variances)

Abbreviation key: EM = expectation-maximization algorithm, **GLM** = generalized linear models.

INTRODUCTION

In the longitudinal data framework, part of the heterogeneity of variances across time in the population can be modeled via random regression (6, 16) or covariance functions (7, 11). Nevertheless, heterogeneity usually remains in the residual variances. More specifically, in the case of the analysis of test-day records for milk production in dairy cattle, different studies (1, 18) have shown that the residual variance changes over time. To cope with this heterogeneity, authors divide the lactation length into different intervals, assuming homogeneity within intervals and heterogeneity between them (6, 13). However, this method can lead to many variance parameters being estimated. This method requires the definition of arbitrary subclasses within which the variance is assumed constant, whereas the change of the residual variance is continuous over time.

Recently, Rekaya et al. (12) proposed a change-point technique to account for the heterogeneity of residual variances along lactation. This approach offers a way to continuously model the changes of the residual variance over time, but assumptions need to be made about the number of change points and the relationship between the residual variance and the number of DIM. Moreover, the number of parameters that have to be estimated may still be quite large and the estimation (using, for instance, Bayesian techniques) time consuming.

The aim of this paper was to propose another way to account for this heterogeneity and to model the changes of the residual variance along lactation as a continuous function of time. For this purpose, a structural model, as proposed by Foulley and Quaas (4) is assumed on the residual variances, and the covariates of this model are parametric functions of time. This procedure offers two main advantages: the number of parameters to be estimated for the residual variances is reduced compared with a purely heterogeneous model, and the changes in the residual variance are considered to be continuous over time, so there is no need to define arbitrary classes of heterogeneity.

The estimates of the parameters for this model on the variances were obtained using an expectation-maximization (**EM**) REML-type algorithm. The equivalence between this system of equations and the generalized linear models (**GLM**) estimating equations has been shown by Lee and Nelder (9). This methodology is illustrated by an analysis of a real data set of monthly records for milk production in dairy cattle.

MATERIALS AND METHODS

Model

Consider a population with I individuals, with individual i having n_i observations. The time and the num-

Received September 27, 1999.
Accepted January 14, 2000.
Corresponding author: F. Jaffrezic; e-mail: florence.jaffrezic@ed.ac.uk.

ber of measurements may be different for each individual. For the sake of simplicity, a simple mixed model (8) for the analysis of longitudinal data was assumed:

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{u}_i + e_{ij} \quad [1]$$

where y_{ij} is measurement j on individual i at time t_{ij} (i = 1, ..., I and j = 1, ..., n_i). $\boldsymbol{\beta}$ is vector of the fixed effects associated with the incidence matrix $\mathbf{X}$ (of row $\mathbf{x}'_{ij}$), and $\mathbf{u}_i$ is the vector of random effects for individual i with incidence matrix $\mathbf{Z}$ (of row $\mathbf{z}'_{ij}$). It is assumed that $\mathbf{u} = (\mathbf{u}'_1, \ldots, \mathbf{u}'_I)' \sim N(0, \mathbf{G})$, and that the residuals e_{ij} are independent, such that

$$e_{ij} \sim N(0, \sigma^2_{e_{ij}}) \quad [2]$$

To model the heterogeneity of the residual variances over time, a structural model (4, 5) was assumed:

$$\ln\sigma^2_{e_{ij}} = \mathbf{p}'_{ij}\boldsymbol{\delta} \quad [3]$$

For instance, if a quadratic function of time is appropriate for the data studied, then

$$\ln\sigma^2_{e_{ij}} = a + bt_{ij} + ct^2_{ij} \quad [4]$$

and $\mathbf{p}'_{ij} = (1\ t_{ij}\ t^2_{ij})$. The model can easily be extended to higher-order polynomials or other parametric functions of time. A stepwise procedure could be used to choose the covariates in the structural model, as discussed by Foulley and Quaas (4).

By using an EM-REML procedure (2) and Lee and Nelder's (9) result (as detailed in the appendix), estimation of all the parameters in this model can be obtained by iterating between the following procedures, which can be achieved with existing software (SAS, Genstat, AS-REML, etc.).

1. Mixed-model equations are constructed assuming a fixed residual variance to obtain estimates of the factors in the model and residuals $\hat{e}_{ij}$.
2. A regression model is applied to the natural log of the squared residuals (GLM equations described in the appendix) to obtain an estimate of $\boldsymbol{\delta}$ in Equation [3].
3. Mixed-model equations are constructed again but using the regression function to determine the appropriate residual variance for each time t_{ij}, and $\sigma^2_{e_{ij}}$, the inverse of which is used as the weighting of the mixed-model equations.
4. Repeat from Step 2 until convergence is reached.

Other algorithms, such as those proposed by Foulley et al. (3), Verbyla (17), and Schnyder et al. (14), could be used for estimating the parameters in the structural model and may differ in convergence rate, ability to remain in the parameter space, and computing time.

Application

The preceding theory was applied to the data set used by White et al. (18). Lactation curves were fitted to test-day records of milk production for 2885 progeny of 30 Holstein-Friesian sires in 503 herds. The lactation stage of animals entering the first test varied between 4 and 40 d, with successive tests at approximate 30-d intervals (10 tests for each cow). The fixed effects considered were the age at calving, the percentage of Holstein genes, and herd-test-month. White et al. (18) considered a sire model and modeled the mean curve of the population as well as the genetic and environmental effects non-parametrically by using smoothing splines.

Here, the exponential curve of Wilmink (19) was fitted as a fixed-regression model for the general mean curve of the population:

$$g(t) = b_0 + b_1 t + b_2 \exp(-Dt) \quad [5]$$

where t is DIM. The parameter D was assumed to be known and equal to 0.068, chosen based on previous studies (1, 18). A sire model was considered, and quadratic random regressions were assumed to model both the genetic and environmental effects:

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + a_{k0} + a_{k1}t_{ij} + a_{k2}t^2_{ij} + b_{i0} + b_{i1}t_{ij} + b_{i2}t^2_{ij} + e_{ij} \quad [6]$$

where y_{ij} is the milk production of cow i taken at time t_{ij}, $\mathbf{x}'_{ij}\boldsymbol{\beta}$ are the fixed effects described above, $a_{k0} + a_{k1}t_{ij} + a_{k2}t^2_{ij}$ is the quadratic random regression for the genetic effect (sire k), and $b_{i0} + b_{i1}t_{ij} + b_{i2}t^2_{ij}$ is the quadratic random regression for the environmental effect (for cow i within sire k). Parameters $\mathbf{a_k} = (a_{k0}, a_{k1}, a_{k2})'$ and $\mathbf{b_i} = (b_{i0}, b_{i1}, b_{i2})'$ are assumed to follow multivariate normal distributions, and e_{ij} is the residual term ($e_{ij} \sim N(0, \sigma^2_{e_{ij}})$).

Two different models for the residual variances were considered: Model 1: 10 classes were assumed for the residual variances, i.e., one for each measurement as considered by White et al. (18); and Model 2: a structural model was assumed on the residual variances, as Equation [4], a quadratic polynomial of time, was being considered. The estimates of the parameters for the latter model were obtained by iterating between AS-REML for the mixed-model equations and SAS for the GLM equations, but this procedure could also easily be incorporated in a REML package.

Table 1. Mean DIM, variance estimates (kg^2), and heritabilities by test.[1]

		Model 1				Model 2			
Test	DIM	G	E	R	h^2	G	E	R	h^2
1	18	3.08	9.17	4.93	0.21	3.11	9.20	5.16	0.21
2	48	3.02	7.90	4.10	0.24	3.05	7.91	4.17	0.24
3	78	3.11	7.42	3.74	0.26	3.13	7.42	3.49	0.27
4	109	3.25	7.31	3.23	0.29	3.26	7.31	3.01	0.29
5	139	3.36	7.33	2.36	0.32	3.37	7.32	2.72	0.31
6	169	3.41	7.33	2.69	0.31	3.42	7.32	2.54	0.32
7	199	3.43	7.33	2.49	0.32	3.44	7.32	2.46	0.32
8	229	3.45	7.49	2.43	0.32	3.47	7.49	2.47	0.32
9	259	3.56	8.09	2.07	0.32	3.60	8.11	2.57	0.31
10	290	3.90	9.63	3.56	0.28	3.96	9.69	2.78	0.29

[1]G = genetic, E = environmental, and R = residual.

RESULTS

Fixed-effect solutions were very similar for both models and were also similar to those obtained by White et al. (18), who fitted a 10-knot spline on the same data set. The breed difference (Holstein-Friesian) was estimated in the first model at 1.56 kg (SE = 0.44) and in the second model at 1.51 kg (SE = 0.44), and the effect of age at calving was 0.18 kg/mo (SE = 0.02) in the two models. As shown in Table 1, the estimates of genetic parameters were also very similar in the two models and were very close to the results of White et al. (18).

Figure 1 shows that the quadratic function was a good representation for the changes of the residual variance. Table 2 gives the estimates of the parameters of the structural model (Model 2) with their standard errors. All were significantly different from 0, and the quadratic function for the residual variances was

$$\ln\sigma^2_{e_{ij}} = 0.97 - 0.073\ t_{ij} + 0.018\ t^2_{ij} \quad [7]$$

where t_{ij} = (DIM − 150)/30. Although this quadratic function seemed to be quite appropriate, the likelihood was higher for the first than the second model (difference of 32 for the log likelihood). Nevertheless, because fewer parameters for the residual variance were estimated in the second (three) than in the first model (ten), a criterion such as Schwarz Bayesian Criterion (15) is more appropriate. This criterion penalizes the likelihood with respect to the number of parameters and is defined by

$$\text{log likelihood} - \frac{1}{2} \times \text{number of parameters in the model} \times \text{Log } n^*$$

where $n^* = n - p$ when using REML with n, the number of observations in the data set, and p, the number of fixed effects. This criterion showed a slightly better fit for the second than the first model (difference of four).

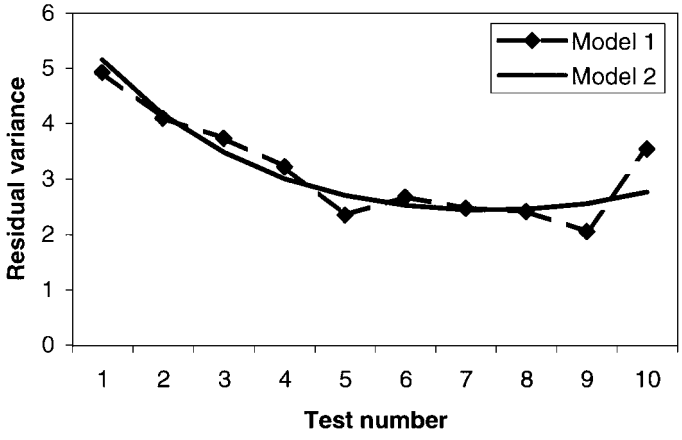

Figure 1. Changes of the residual variance over time for the two models. Model 1: 10 different classes of heterogeneity (1 for each test); Model 2: structural model on the residual variance ($\ln\sigma^2_{e_{ij}} = 0.97 - 0.073\ t_{ij} + 0.018\ t^2_{ij}$, where t_{ij} = (DIM − 150)/30).

DISCUSSION

The improvement in fit of the structural model on the residual variance compared with the heterogeneous model (assuming 10 different classes of heterogeneity) was not great. However this method would prove to be much better in the case of high heterogeneity within classes. For instance, with Model 1 modified so that the

Table 2. Estimates and SE of the parameters of the structural model on the residual variances (Model 2) ($P < 0.001$ for each parameter).

Parameters	Estimate	SE
a	0.970	0.008
b	–0.073	0.002
c	0.018	0.001

lactation was divided into 5 intervals rather than 10, the likelihood was greater for the second than the first model (difference of nine for the log likelihood), even though Model 2 still had fewer parameters.

Nevertheless, the polynomial functions may not be the most appropriate, especially because of their lack of flexibility to model the variances at the beginning and at the end of the lactation. Other more flexible parametric functions could be considered using the same methodology.

This method offers two important advantages: fewer parameters are estimated than in the classical heterogeneous model, and the variance is a continuous function of time with no arbitrary classes. This approach could also be a useful alternative for other longitudinal studies that arise in animal breeding, for instance growth curve analyses.

Other factors of heterogeneity could be taken into account in the structural model on the residual variances (4), for instance the age at calving, month of calving, region, year, and even the herd-test-month (perhaps as a random effect). This aspect of the heterogeneity of variances, which applies to the residual variances as well as the genetic and permanent environmental variances, should be investigated more thoroughly.

ACKNOWLEDGMENTS

We are most grateful to Sue Brotherstone for her contribution to the application on dairy cattle data. Thanks also to Peter Visscher, Vincent Ducrocq, Jean-Louis Foulley, Christèle Robert-Granié, R. L. Quaas, and an anonymous reviewer for helpful comments and ideas. This work was supported by the Department of Animal Genetics of the INRA (National Institute of Agronomical Research), Jouy-en-Josas, France.

REFERENCES

1 Brotherstone, S., I.M.S. White, and K. Meyer. Genetic modelling of daily milk yield using orthogonal polynomials and parametric curves. Anim. Sci. (in press).
2 Dempster, A., N. Laird, and R. Rubin. 1977. Maximum likelihood estimation from incomplete data via the EM algorithm. J. R. Stat. Soc. 39:1–20.
3 Foulley, J. L., D. Gianola, M. San Cristobal, and S. Im. 1990. A method for assessing extent and sources of heterogeneity of residual variances in mixed linear models. J. Dairy Sci. 73:1612–1624.
4 Foulley, J. L., and R. L. Quaas. 1995. Heterogeneous variances in gaussian linear mixed models. Genet. Sel. Evol. 27:211–228.
5 Foulley, J. L., R. L. Quaas, and C. Thaon d'Arnoldi. 1998. A link function approach to heterogeneous variance components. Genet. Sel. Evol. 30:27–43.
6 Jamrozik, J., and L. Schaeffer. 1997. Estimates of genetic parameters for a test day model with random regressions for yield traits of first lactation Holsteins. J. Dairy Sci. 80:762–770.
7 Kirkpatrick, M., W. G. Hill, and R. Thompson. 1994. Estimating the covariance structure of traits during gorwth and ageing: illustrated with lactation in dairy cattle. Genet. Res. 64:57–69.
8 Laird, N. M., and J. H. Ware. 1982. Random effects models for longitudinal data. Biometrics 38:963–974.
9 Lee, Y., and J. A. Nelder. 1999. Extended REML using GLM technology: a new formulation. Tech. Rep. Dep. Math., Imperial College, London, United Kingdom.
10 McCullagh, P., and J. A. Nelder. 1989. Generalized Linear Models. 2nd ed. Chapman and Hall, London, United Kingdom.
11 Meyer, K., and W. G. Hill. 1997. Estimation of genetic and phenotypic covariance functions for longitudinal or repeated records by Restricted Maximum Likelihood. Livest. Prod. Sci. 47:185–200.
12 Rekaya, R., M. J. Carabano, and M. A. Toro. 1998. Assessment of heterogeneity of residual variances using changepoint techniques. 49th Annu. Mtg. Eur. Assoc. Anim. Prod., Warsaw, Poland.
13 Rekaya, R., M. J. Carabano, and M. A. Toro. 1999. Use of test day yields for the genetic evaluation of production traits in Holstein-Friesian cattle. Livest. Prod. Sci. 57:203–217.
14 Schnyder, U., A. Hofer, F. Labroue, and N. Kunzi. 1999. Genetic parameters of a random regression model for daily feed intake of performance tested French Landrace and Large White growing pigs. 50th Annu. Mtg. Eur. Assoc. Anim. Prod., Zurich, Switzerland.
15 Schwarz, G. 1978. Estimating the dimension of a model. Ann. Stat. 6:461–464.
16 Verbeke, G., and G. Molenberghs. 1997. Linear Mixed Models in Practice: a SAS oriented approach. Lecture Notes in Statistics 126. Springer-Verlag, New York, NY.
17 Verbyla, A. P. 1993. Modelling variance heterogeneity: residual maximum likelihood and diagnostics. J. R. Stat. Soc. 55:493–508.
18 White, I.M.S., R. Thompson, and S. Brotherstone. 1999. Genetic and environmental smoothing of lactation curves with cubic splines. J. Dairy Sci. 82:632–638.
19 Wilmink, J.B.M. 1987. Adjustment of test day milk, fat and protein yield for age, season and stage of lactation. Livest. Prod. Sci. 16:335–348.

APPENDIX

The REML estimates of the parameters in the structural model for the residual variance were obtained using an EM algorithm (2).

Letting $\mathbf{c} = (\mathbf{y}', \theta')'$ be the complete set of data, and $\Theta = (\beta', \mathbf{u}')'$ the vector of the missing values. The likelihood function of the complete data is

$$p(\mathbf{c}|\delta, \mathbf{G}) = p(\mathbf{y}|\beta,\mathbf{u}, \delta)p(\beta, \mathbf{u}|\mathbf{G}) \quad [8]$$

Therefore, the log likelihood is

$$-2\ln p(\mathbf{c}|\delta, \mathbf{G}) = -2L(\delta, \mathbf{G}; \mathbf{c}) = -2L(\delta; \mathbf{e}) - 2L(\mathbf{G}; \mathbf{u}) \quad [9]$$

and the estimation of δ can then be separated from that of $\mathbf{G}$, considering the log likelihood

$$-2L(\delta; \mathbf{e}) = \text{const.} + \sum_{i=1}^{I}\sum_{j=1}^{n_i}[\ln\sigma^2_{e_{ij}} + \frac{1}{\sigma^2_{e_{ij}}}e^2_{ij}] \quad [10]$$

The E-step is defined as usual, i.e., at iteration (r) one calculates the conditional expectation of $L(\delta; \mathbf{e})$ given the data $\mathbf{y}$ and $\delta = \delta^{(r)}$.

$$Q(\delta|\delta^{(r)}) = E(-2L(\delta;\ \mathbf{e})|\mathbf{y},\ \delta^{(r)}) \quad [11]$$

$$Q(\delta|\delta^{(r)}) = \text{const.} + \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left[\ln\sigma^2_{e_{ij}} + \frac{1}{\sigma^2_{e_{ij}}}E_c(e^2_{ij})\right] \quad [12]$$

where $E_c(e^2_{ij})$ is the conditional expectation $E(e^2_{ij}|\mathbf{y},\ \delta^{(r)})$, and

$$E(e^2_{ij}|\mathbf{y},\ \delta^{(r)}) = (E(e_{ij}|\mathbf{y},\ \delta^{(r)}))^2 + \text{trace}(\text{Var}(e_{ij}|\mathbf{y},\ \delta^{(r)})) \quad [13]$$

$$= \hat{e}^2_{ij} + \text{Var}(e_{ij}|\mathbf{y},\ \delta^{(r)}) \quad [14]$$

The M-step consists of calculating the next value $\delta^{(r+1)}$ by minimizing the function $Q(\delta|\delta^{(r)})$with respect to δ,

$$\frac{\partial Q}{\partial \delta} = \frac{\partial Q}{\partial \sigma^2_{e_{ij}}}\frac{\partial \sigma^2_{e_{ij}}}{\partial \ln\sigma^2_{e_{ij}}}\frac{\partial \ln\sigma^2_{e_{ij}}}{\partial \delta}. \quad [15]$$

Then

$$\frac{\partial Q}{\partial \delta} = \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left[1 - \frac{1}{\sigma^2_{e_{ij}}}E_c(e^2_{ij})\right]\mathbf{p}_{ij} \quad [16]$$

$$\frac{\partial Q}{\partial \delta} = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(1 - w_{ij})\left(1 - \frac{1}{\sigma^2_{e_{ij}}}d^*_{ij}\right)\mathbf{p}_{ij} \quad [17]$$

where $w_{ij} = \frac{1}{\sigma^2_{e_{ij}}}\text{Var}(e_{ij}|\mathbf{y},\ \delta^{(r)})$ and $d^*_{ij} = \hat{e}^2_{ij}/(1 - w_{ij})$.

Lee and Nelder (9) showed that this system of equations is equivalent to the estimating equations for a GLM (10) with response d^*_{ij} [where d^*_{ij} is the square of the residuals divided by the weight $(1 - w_{ij})$], mean $\sigma^2_{e_{ij}}$, error gamma, log-link $(\ln(\sigma^2_{e_{ij}}))$, linear predictor $\xi_{ij} = \mathbf{p}'_{ij}\delta$, and prior weight $(1 - w_{ij})$.

The values of $\hat{e}^2_{ij}$ and $\text{Var}(e_{ij}|\mathbf{y},\ \delta^{(r)})$ can be calculated from the solutions of the mixed-model equations as follows:

$$\hat{e}_{ij} = y_{ij} - \mathbf{x}'_{ij}\hat{\beta} - \mathbf{z}'_{ij}\hat{\mathbf{u}} \quad [18]$$

where $\hat{\beta}$ and $\hat{\mathbf{u}}$ are the BLUP solutions.

Letting $\theta = (\beta', \mathbf{u}')'$ and $\mathbf{b}_{ij} = (\mathbf{x}'_{ij},\ \mathbf{z}'_{ij})$, then $e_{ij} = y_{ij} - \mathbf{b}_{ij}\Theta$ and, therefore,

$$\text{Var}(e_{ij}|\mathbf{y},\ \delta^{(r)}) = \mathbf{b}_{ij}\ \text{Var}(\Theta|\mathbf{y},\ \delta^{(r)})\mathbf{b}'_{ij} \quad [19]$$

where $\text{Var}(\theta|\mathbf{y},\ \delta^{(r)})$ corresponds to the inverse of the coefficient matrix in the mixed-model equations.