MODELLING OF LONGITUDINAL LIVEWEIGHT DATA USING REGRESSION WITH LEGENDRE AND EIGENVECTOR FUNCTIONS

M. D. Price, M. E. Olayemi and J. B. Bryant

DairyNZ Limited, Private Bag 3221, Hamilton 3240, New Zealand

SUMMARY

Liveweight data from walk-over-weigh technology on NZ dairy farms provides a potential source of daily or average liveweight data for genetic selection. In this study, mixed models utilising both fixed and random regression with varying orders of Legendre polynomials or eigenvectors were assessed to model the longitudinal properties of liveweight records of New Zealand dairy cows. Higher order models fitted the data best despite the subsequently large increase in number of parameters. The choice of either Legendre polynomials or eigenvectors as the set of basis functions made no significant difference. Limits to model order will need to be applied on the basis of assumptions concerning data covariance structure and desired number of liveweight-derived traits for selection. Additionally, methods developed here for estimation of genetic covariance of regression coefficients and for inversion of the mixed model equation matrix can be extended to models for which pedigree relationships are also taken into account.

INTRODUCTION

Liveweight (LW) is a trait which changes over time, and may thus be considered as "longitudinal" or "infinite-dimensional" (Kirkpatrick *et al.*, 1994). Some of the approaches that have been applied to model these repeated data types include repeatability, multivariate and random regression (RR) models (Mrode and Thompson, 2014).

The use of RR models has become a preferred method to analyse longitudinal data for animal genetic evaluation. These models use a fixed regression to describe the average shape of a lactation or growth curve, and a random regression for each animal to account for deviations from the fixed regression (Schaeffer and Deckers, 1994). RR models have been applied to model test-day records of milk traits in dairy cattle (Jamrozik and Schaeffer, 1997; Olori *et al.*, 1999), as well as growth and mature weight in beef cattle (Meyer, 1999, 2004; Speidel *et al.*, 2010) and other livestock. In these models, Legendre polynomials are typically used as the set of basis functions, however they can suffer from a Runge effect (Runge, 1901), where a higher order polynomial describing the general curve has high oscillations in the boundary areas. Eigenvectors avoid this problem, and just a few eigenvectors may adequately account for the covariance structure of the data, so they may provide a better set of basis functions in a RR model.

The current animal evaluation model for LW of NZ dairy cattle is based on a combination of both visual score and static weights. Walk-over-weigh (WOW) records per animal in commercial dairy sheds provide the opportunity to incorporate this new data into the current LW models. Utilisation of these data may also allow for better characterisation of the seasonal LW curve. The objective of this study was to compare RR models using either Legendre or eigenvector basis functions of various orders with longitudinal LW data, albeit without considering pedigree.

MATERIALS AND METHODS

Data. A total of N = 58,532 WOW records collected on n_a = 2,899 2-year-old New Zealand dairy cows born in the 2010/2011 season were extracted from the Dairy Industry Good Animal Database (DIGAD). The data consisted of individual animal weekly average LW collected over n_t = 40 weeks from lactation start (defined as weeks-in-milk, WIM), from n_{cg} = 6 herds. All data manipulation, modelling and analysis were performed with R statistical software.

Models. Mixed models were used in the analysis, whose fixed component included a regression on either Legendre or eigenvector basis functions which would describe the general liveweight curve, and whose random component described individual deviations from the general curve. Following the notation of Mrode & Thompson (2014), models used were of the form:

$$y_{tij} = htd_i + \sum_{k=0}^{n_f} \psi_{tk}\beta_k + \sum_{k=0}^{n_r} \phi_{tk}a_{jk} + e_{tij} \quad (1)$$

Where y_{tij} is the test day record for cow j on day t within contemporary group (herd test day) i, ψ_{tk} is the value of a k^{th} basis function (Legendre or eigenvector) evaluated at time t, $\beta_k \in \mathbf{\beta}$, the vector of fixed regression coefficients, ϕ_{tk} is the value of a k^{th} Legendre polynomial evaluated at time t, and $a_{jk} \in \mathbf{a}_j$, the vector of random regression coefficients (animal effects) for animal j (where $\mathbf{a}_j \in \mathbf{a}$, the full vector of random regression components). The set of basis functions for fixed effects are of order n_f, and for random effects are of order n_r. The matrix notation (2) and mixed model equation (MME) notation (3) for this model are as follows:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e} \quad (2)$$

$$\begin{bmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{X} & \mathbf{X}^{\mathrm{T}}\mathbf{Z} \\ \mathbf{Z}^{\mathrm{T}}\mathbf{X} & \mathbf{Z}^{\mathrm{T}}\mathbf{Z} + \mathbf{I}\sigma_e^2 \otimes \mathbf{K}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{y} \\ \mathbf{Z}^{\mathrm{T}}\mathbf{y} \end{bmatrix} \quad (3)$$

Here it is assumed that var(**a**) = $\mathbf{I}\otimes\mathbf{K}$ and var(**e**) = $\mathbf{I}\sigma_e^2$ (necessary priors), where $\otimes$ is the Kronecker product and **K** is an n_r-dimensional covariance structure of the random regression coefficients for animal effects. **X** and **Z** are the incidence matrices corresponding to the effect solutions (superscript T indicates matrix transpose). Here **Xb** = $\mathbf{X}_1\mathbf{b}_1+\mathbf{X}_2\mathbf{\beta}$, the sum of contemporary group fixed effects ($htd_i \in \mathbf{b}_1$) and fixed regression components respectively. It should be noted that our approach does not yet take into account the pedigree relationships, which would otherwise partition the animal effect into additive animal genetic and permanent environmental effects. If this were the case, the MME would be in a 3×3 partition, with n_r-dimensional estimated covariance structures of **G** and **P**, in Kronecker product with $\mathbf{A}^{-1}\sigma_e^2$ and $\mathbf{I}\sigma_e^2$, respectively. As it stands, we instead used an estimate of the combined covariance of **G** and **P**; **K**, which was the genetic covariance of animal effects $\mathrm{cov_A}(\mathbf{a})$ calculated from the phenotypic covariance of the data $\mathrm{cov_P}(\mathbf{y})$, with **y** structured as an animal × test-week dataset. Liveweight heritability was assumed to be $h^2 = 0.35$; residual variance was assumed to be $\sigma_e^2 = 400$kg.

$$\mathbf{K} = \mathrm{cov_A}(\mathbf{a}) = \frac{h^2}{2}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathrm{cov_P}(\mathbf{y})((\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}})^{\mathrm{T}} \quad (4)$$

Basis function sets. For Legendre polynomials, the setup of an order k incidence matrix (for a basis set of k functions for either the fixed or random model component) was calculated as **X** (or **Z**) = **ML**, where **L** was the $k\times k$ matrix of Legendre polynomial coefficients, and **M** was the $N\times k$ matrix of each observation's week-in-milk t transformed to the [-1, 1] interval (5) and evaluated for the k different degrees of monomials.

$$x_m = \frac{2(t_m - t_{min})}{(t_{max} - t_{min})} - 1, \qquad m \in \{1 \cdots N\} \quad (5)$$

For eigenvectors, the setup of an order k incidence matrix was calculated as $\mathbf{X} = \mathbf{TE}_{\mathrm{k}}$, where **T** was the $N\times n_t$ matrix for each observation's WIM, and $\mathbf{E}_{\mathrm{k}}$ was the $n_t\times k$ subset of the k eigenvectors of the top k eigenvalues of the eigendecomposition of the phenotypic covariance matrix, $\mathrm{cov_P}(\mathbf{y}) = \mathbf{E\Lambda E}^{\mathrm{T}}$ (**E** is the matrix of all eigenvectors; $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues).

Residual analysis. Akaike information criteria (AIC) (Akaike, 1974) were calculated for each model to provide a relative measure of model quality. For nested models, Likelihood ratio tests (LRT) (proven by the Neyman–Pearson (1933) lemma to be optimal for model selection) were also used to determine if any reduction in residuals between models was significant or not.

Assuming normal distribution of residuals, the log-likelihood of a model was determined as a function of residual variance σ^2 and number of observations N:

$$\ell n(\mathcal{L}) = -\frac{N}{2}\left(1 + \ell n(2\pi\sigma^2)\right) \qquad (6)$$

From this, AIC = $2n$ - $2\ell(\mathcal{L})$ was calculated (where $n = n_{cg} + n_f + n_r .n_a$ was the total number of parameters), and for two nested models a likelihood ratio LR = $2\ell n(\mathcal{L}_{M2})$ - $2\ell n(\mathcal{L}_{M1})$ was calculated (for "null" model M1 nested within model M2). A chi-squared test using test statistic LR with degrees of freedom $df = n_{M2} - n_{M1}$ would produce a value $p = 1 - \chi^2$ which, for $p < 0.05$, would indicate that model M2 was significantly better than the simpler M1 model.

MME matrix inversion. The block-structure of the MME matrix (dimension $n_{cg} + n_f + n_r.n_a$) allowed for an alternative inversion of much smaller n_r-dimensional matrices. If the data is ordered by animal, then the $\mathbf{Z}^T\mathbf{Z}$ component of the MME (3) will be block-diagonal, comprised of n_a sub-matrices of dimension n_r each. The same is true for $\mathbf{I}\sigma_e^2 \otimes \mathbf{K}^{-1}$, and so $\mathbf{D} = \mathbf{Z}^T\mathbf{Z} + \mathbf{I}\sigma_e^2 \otimes \mathbf{K}^{-1}$ will be block-diagonal also. Given the Banachiewicz (1937) identity for of a partitioned matrix inverse,

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{S}_D^{-1} & -\mathbf{S}_D^{-1}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{S}_D^{-1} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{S}_D^{-1}\mathbf{B}\mathbf{D}^{-1} \end{bmatrix}, \quad \text{where } \mathbf{S}_D = \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C} \qquad (7)$$

and the property of the inverse of a block-diagonal matrix being another block-diagonal matrix of individual block inverses, then it follows that the inverse of the MME matrix may simply be determined by way of calculating the inverse of each n_r-dimensional sub-matrix of $\mathbf{D}$ (and the inverse of the small (n_{cg}+n_f)-dimensional Schur matrix $\mathbf{S}_D$).

RESULTS AND DISCUSSION

Varying model orders. For RR models with fixed Legendre component of $n_f \in \{3, \ldots, 7\}$ and random component of $n_r = 3$, the general polynomial representing the data (Figure 1) shows how the oscillations of higher order polynomials become apparent for $n_f \gtrsim 5$. Despite this, LRT's between models of consecutive orders of fixed Legendre component $n_f \in \{3, \ldots, 20\}$ with conserved random order $n_r \in \{1, \ldots, 3\}$ showed significantly better fit for orders $n_f \leq 9$ ($p < 0.005$). However, residual variance was not longitudinally homogenous (even for varying random orders), indicating that higher order fixed effects may be required. Unlike the random regression component, increasing the fixed order makes almost no computational difference. However, under the assumption of a reasonably smooth longitudinal relationship between LW values, the fixed order should be limited.

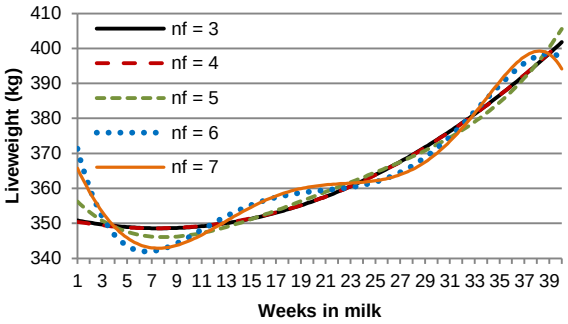

Figure 1. Fixed regression curves for models with Legendre orders $n_f \in \{3, \ldots, 7\}$ and $n_r = 3$.

For RR models with fixed Legendre component of $n_f = 9$ and random component of $n_r \in \{1, \ldots, 5\}$, residual variance decreases (and log-likelihood increases) with increasing n_r (Table 1), and this variance also scales quite uniformly across season for different n_r. LRT's indicated that models were very significantly better with increase in random component order for $n_r \leq 3$, for both Legendre and eigenvectors models. The AIC was minimised for $n_r = 3$, indicating that quadratic random animal effects are best. Increased parameters per animal should contribute in a biologically meaningful way, as it is upon those parameters that animals may be selected by. Therefore, while incorporating a random regression component into the model is advisable, the order of this component should not be too large; letting $n_r = 3$ should be sufficient.

Table 1. AIC and LRT's of models of Legendre or eigenvector fixed component $n_f = 9$, and $n_r \in \{1, \ldots, 5\}$

	Legendre, $n_f = 9$					Eigenvector, $n_f = 9$				
Order n_r	1	2	3	4	5	1	2	3	4	5
$\ell n(\mathcal{L})$ ($\times 10^3$)	-289	-280	-276	-273	-271	-289	-280	-276	-274	-272
AIC ($+5.7\times 10^5$)	13,904	2,533	449	1,051	2,895	14,661	2,930	568	1,217	3,256
LR (n_r vs. n_{r-1})		17,169	7,882	5,196	3,954		17,529	8,160	5,148	3,760
$p = 1\text{-}\chi^2$		0	0	1	1		0	0	1	1

Legendre vs. eigenvector. Models with a fixed regression component of either Legendre polynomials or eigenvectors of order $n_f = 9$ were compared via AIC values for their relative merit, for random component of Legendre orders $n_r \in \{1, \ldots, 5\}$. For any particular random order n_r, Legendre models had slightly better AIC values than those of the eigenvector models (Table 1). In the absence of a measure of significance for AIC comparison, the default choice of regression function ought to remain as the Legendre polynomials.

Pedigree information. Future inclusion of pedigree information ought to improve model fit even more due to increased utilisation of data via pedigree linkages. While the relationship matrix **A** would not be subject to inversion by the block-matrix inversion technique, the use of $\mathbf{A}^{-1}$ would allow for a similar technique for solving the model. Therefore the use of a full (including pedigree) RR model will have essentially no more computational complexity for random orders $n_r > 1$.

CONCLUSION

In a RR model for WOW data, increasing orders of fixed and random regression components significantly improve the model in general, though these must be tempered by the practical realities of assumed longitudinal relationships and necessary number of parameters per animal. The choice of type of regression function (Legendre polynomials or eigenvectors) is insignificant.

The future inclusion of pedigree relationships ought to ensure a much better depth of data per animal and subsequent model improvement. The use of block-matrix inversion will still ensure that computational complexity is significantly reduced.

REFERENCES

Akaike H. (1974) *IEEE Transactions on Automatic Control* **19**(6): 716.

Banachiewicz T. (1937) *Acta Astronomica, Series C*, **3**: 41.

Jamrozik J. and Schaeffer L. R. (1997) *J. Dairy Sci*. **80**: 762.

Kirkpatrick M., Hill W.G. and Thompson R. (1994) *Genet. Res.* **64**: 57.

Meyer K. (1999) *J. Anim. Bree. Genet.* **116**: 181.

Meyer K. (2004) *Livest. Prod. Sci*. **86**: 69.

Mrode R.A. and Thompson R. (2014) '*Linear Models for the Prediction of Animal Breeding Values', 3rd Ed. Cabi Publ., Wallingford, UK*. pp130–155.

Neyman J. and Pearson E.S. (1933) *Phil. Trans. Royal Soc. A: Math. Phys. Eng. Sci.* **231**(694–706):289.

Olori V.E., Hill W.G., McGuirk B.J. and Brotherstone S. (1999) *Livest. Prod. Sci*. **61**: 53.

R Core Team (2014). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria*. URL http://www.R-project.org/

Runge C. (1901) *Zeitschrift für Mathematik und Physik* **46**: 224.

Schaeffer L.R. and Dekkers J.C.M. (1994) *Proc. 5th World Congr. Genet. Appl. Livest. Prod., Guelph, Canada* **18**: 443.

Speidel S.E., Enns R.M. and Crews D.H. Jr. (2010) *Genet. Mol. Res*. **9**: 19.