Proceedings of the Beef Improvement Federation 41st Annual Research Symposium
April 30 – May 3, 2009, Sacramento, California, USA

Genetic Analysis of Longitudinal Data in Beef Cattle

Scott Speidel, Colorado State University

Introduction

In today's beef industry, many different data types are collected by beef cattle breed associations for the purpose of genetic evaluation. These data points are all biological characteristics of individual animals which can be measured a multitude of times over an animal's lifetime. The number of times a given trait is observed during an animal's life is dictated by the nature of the trait. For example, traits such as carcass characteristics, heifer calving ease, and heifer pregnancy can only be recorded one time on an individual animal. However, traits which monitor the status of an animal as it grows such as weight traits and live animal indicators of carcass merit can be measured a number of times over the lifespan of an animal. Weight traits such as birth weight and weaning weight describe the same underlying trait, growth as measured by weight gain observed over time. As such, perhaps they can be best described by some type of mathematical function rather than a finite set of data points (Kirkpatrick and Heckman, 1989; Meyer and Kirkpatrick, 2005). As a result, this unique data type has been referred to throughout the literature as "function valued" (Kirkpatrick and Heckman, 1989; Meyer and Kirkpatrick, 2005) or as "infinite-dimensional" or "longitudinal" data by Meyer and Hill (1997). This essay summarizes past and current genetic evaluation technology for handling longitudinal data, and reviews some emerging technologies that are beginning to be implemented in current national cattle evaluation schemes.

Literature Review

Longitudinal Data. A number of traits currently collected for beef cattle genetic evaluation fall under the umbrella definition of longitudinal data. They can range from commonly collected observations such as weight, height and body condition score measurements to more obscure measures such as feed intake, survival and sperm production and quality (Schaeffer, 2004). Several different methods have been implemented by groups conducting national cattle evaluations to properly model these

data types. These methods include more traditional models such as the repeatability and multivariate models, to the more contemporary (and perhaps more appropriate) models such as the suite of random regression models using different base functions (Mrode, 2005).

The analysis of function valued traits is challenging, and each of the different methods has their respective benefits and limitations. Discussion of these benefits and limitations for each of the methods of analysis will be addressed individually beginning with the traditional repeatability model, then move on to the multivariate models and finally finishing with random regression models that use covariance functions and splines as their base function.

The Repeatability Model. Perhaps the simplest method of analysis of longitudinal data is the "Repeatability Model". The idea behind this model is to treat each observation as a repeated record of the same trait on the same individual. This model has been implemented in the past for traits such as litter size in successive pregnancies in swine and milk yield in successive lactations (Jamrozik et al., 1997b; Interbull, 2000).

The repeatability model is most often described in matrix form by the following (Mrode 2005):

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{Wp} + \mathbf{e},$$

where **X**, **Z**, and **W** are incidence matrices relating the repeated observations in **y** to fixed (**b**), random additive animal genetic (**u**), and random permanent environmental and non-additive genetic effects (**p**), with **e** defining a vector of random residual errors. The model makes the assumption that the mean of the random effects is zero with variances represented by:

$$\text{var}\begin{bmatrix}\mathbf{u}\\ \mathbf{p}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{A}\sigma_u^2 & 0 & 0\\ 0 & \mathbf{I}\sigma_p^2 & 0\\ 0 & 0 & \mathbf{I}\sigma_e^2\end{bmatrix},$$

where σ_u^2, σ_p^2, and σ_e^2 are the variances of random additive animal genetic effect, random permanent environmental effect, and random residual error, respectively. In the above **A** is Wright's numerator relationship matrix (Wright, 1922) and **I** is an identity matrix with an order equal to the number of observations in **y**. The observations in **y**

are assumed to have the mean **Xb** and variance equal to $\text{var}(\mathbf{y}) = \mathbf{ZAZ}'\sigma_a^2 + \mathbf{WI}\sigma_p^2\mathbf{W}' + \mathbf{I}\sigma_e^2$.

As can be inferred from the model presented above, the repeatability model makes assumptions on the data structure that do not hold under all situations. Under the assumptions of the repeatability model, observations from the same individual measured at different ages are assumed to have a constant variance and a common correlation with each other (Jennrick and Schluchter, 1986). This assumption of constant variance does not hold where individual variance changes according to the amount of time that has passed between measurements (Meyer and Hill 1997). In the situation where the repeated observations typically follow some type of curve (e.g. growth or lactation curves) correlations between observations taken close together in time are higher than those taken farther apart from one another. In this situation, a more complex model that accounts for the differing correlation structure between successive observations is required.

The Multiple Trait Model. Multivariate genetic evaluation, introduced by Henderson and Quaas (1976), predicts genetic values for multiple traits through the incorporation of genetic and residual covariances among the traits (Mrode, 2005). This property can be extended to the analysis of longitudinal data if differing measurements on an individual animal are treated as separate but genetically correlated traits. It is under this assumption that the current national cattle genetic evaluations for growth are performed. For example, birth weight and weaning weight are observations which are analyzed as separate but genetically correlated traits using a multivariate model even though both are observations of the growth of an individual.

This multivariate model as described by Mrode (2005) is shown in matrix form below.

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1 & 0\\ 0 & \mathbf{X}_2\end{bmatrix}\begin{bmatrix}\mathbf{b}_1\\ \mathbf{b}_2\end{bmatrix} + \begin{bmatrix}\mathbf{Z}_1 & 0\\ 0 & \mathbf{Z}_2\end{bmatrix}\begin{bmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{bmatrix} + \begin{bmatrix}\mathbf{e}_1\\ \mathbf{e}_2\end{bmatrix}$$

In the above set of equations, $\boldsymbol{y_i}$ is a vector of observations for the i^{th} trait, $\boldsymbol{b_i}$ is a vector of fixed effects for the i^{th} trait, $\boldsymbol{u_i}$ and $\boldsymbol{e_i}$ are vectors of random animal genetic and random residual effects for the i^{th} trait, respectively. $\boldsymbol{X_i}$ and $\boldsymbol{Z_i}$ are incidence matrices

relating the observations in **y** to the fixed effects in **b** and random animal genetic effects in **u**. As with the above repeatability model, the observations in **y** are assumed to have the mean **Xb**. Random effects in the model are assumed to have means of zero and genetic variances equal to:

$$\text{var}\begin{bmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{bmatrix} = \begin{bmatrix}\sigma^2_{g_1} & \sigma_{g_1,g_2}\\ \sigma_{g_2,g_1} & \sigma^2_{g_2}\end{bmatrix} \otimes \mathbf{A}$$

and residual variances equal to:

$$\text{var}\begin{bmatrix}\mathbf{e}_1\\ \mathbf{e}_2\end{bmatrix} = \begin{bmatrix}\mathbf{I}\sigma^2_{e_1} & \mathbf{I}\sigma_{e_1,e_2}\\ \mathbf{I}\sigma_{e_2,e_1} & \mathbf{I}\sigma^2_{e_2}\end{bmatrix}.$$

Above, $\sigma^2_{g_1}$, $\sigma^2_{g_2}$,σ_{g_1,g_2}, and σ_{g_2,g_1} are the additive genetic variance for $\mathbf{y}_1$, $\mathbf{y}_2$ and the additive genetic covariances between $\mathbf{y}_1$ and $\mathbf{y}_2$, respectively. Likewise, $\sigma^2_{e_1}$, $\sigma^2_{e_2}$, σ_{e_1,e_2}, and σ_{e_2,e_1} are the residual error variances for $\mathbf{y}_1$ and $\mathbf{y}_2$ as well as the residual covariances between $\mathbf{y}_1$ and $\mathbf{y}_2$. **A** is Wright's numerator relationship matrix and **I** is an n × n identity matrix.

Henderson and Quaas (1976) were the first to implement the multivariate BLUP model illustrated above in the analysis of a three trait beef cattle example (birth weight, weaning weight and post-weaning gain). Following their work, Schaeffer and Jamrozik (1996) first suggested the use of a multivariate model for the analysis of test day records for milk volume, fat, and protein percentages in dairy cattle. In each of these examples, the observations measured on individuals across time were treated as separate and unique traits that are genetically correlated to one another.

The multivariate model is not without its inherent problems when analyzing longitudinal data. Given the fact longitudinal data can be described using some type of function (Meyer and Kirkpatrick, 2005), they tend to have a large number of data points which are of interest to the individuals performing the data collection. In the multivariate model, this can lead to equation systems which have very high dimension and computational requirements. Considering the test day records discussed by Wiggans and Goddard (1996, 1997) three yield traits (milk volume, fat and protein percentages) over two parity groups (first parity versus later parities) and ten stages of lactation (ten

different test days per lactation), analyzing this data using a multivariate model would result in an analysis with 60 different traits.

Another issue with the multivariate model is the potential for high correlations between successive measurements. In beef cattle evaluation, weaning weight and yearling weight are two traits of economic importance, with genetic and phenotypic correlations between these two measurements reported to be 0.78 and 0.72, respectively (Koots, 1994). In the analysis of test day records, the correlations are even higher. Pander et al., (1992) reported milk yield correlations ranging from 0.97 (1 test day apart) to 0.73 (7 test days apart), with correlations between fat yield and protein yield test day records nearly as high. These elevated correlations are undesirable for two main reasons. First, if two variables predict the same information, it doesn't make sense to include both of the variables in the model. Second, the correlation between the two variables has the effect of reducing the power of the tests of significance (Foster et al., 2006).

The high correlations between traits such as weaning and yearling weights as well as between individual test days in dairy cattle evaluation have resulted in studies designed to determine how to specifically handle these elevated correlations. One method, an extension of the multivariate model, allows higher correlations between observations measured close together than those measured farther apart. This technique, referred to as autoregression or autocorrelation, has been documented in the literature numerous times (Harville, 1979; Kachman and Everett, 1993; Carvalheira et al., 1998). Another method to handle this data type is to model the data using a pre-determined function, or data mean. Referred to as fixed regression (Mrode, 2005), these functions can be extended in such a manner where each individual will have its own random function.

Random Regression. Regression models have been used in the analysis of longitudinal data for many years. The use of pre-determined functions as covariates was introduced as random regression or a random coefficients model during the early to mid 1980's (Henderson, 1982; Laird and Ware, 1982; Jennrich and Schluchter, 1986). However, the first study with application to livestock production data was conducted by Ptak and

Schaeffer (1993) in the analysis of test day milk production records of dairy cattle. This first attempt was not a random regression model, but it accounted for the general shape or mean lactation curve for cows within similar herd, year and season. Following this initial trial, Schaeffer and Dekkers (1994) extended the regression coefficients of this fixed regression model to random animal effects. In doing so, they were able to account for the mean shape of the lactation curve within a given herd, year and season, as well as account for the deviation of each individual animal's lactation curve from this mean shape. They were also able to account for the change in correlation structure of repeated records on individuals over time. This ability of the random regression model to properly account for the changing correlation structure has been shown to result in an increase in prediction accuracy of 5.9% when compared to the multivariate model (Meyer, 2004).

The general form of a random regression model as described by Mrode (2005) can be shown in matrix form as:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Qu} + \mathbf{Zpe} + \mathbf{e}$$

where **y** is a vector of repeated test day yields measured on individual animals, **X** is an incidence matrix relating observations in **y** to fixed effects and fixed regression coefficients, **b** is a vector of solutions for fixed effects and fixed regressions, **Q** is an incidence matrix of covariates relating observations in **y** to random additive genetic regression coefficients, **u** is a vector of random additive direct genetic effects, **Z** is an incidence matrix of covariates relating observations in **y** to permanent environmental random regression coefficients, **pe** is a vector of random permanent environmental regression coefficients for each animal, **e** is a vector of random residuals which includes the temporary environmental effects for each observation. Variances assumed for this model are:

$$\text{var}\begin{bmatrix}\mathbf{u}\\ \mathbf{pe}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{A}\otimes\mathbf{G} & 0 & 0\\ 0 & \mathbf{I}\otimes\mathbf{P} & 0\\ 0 & 0 & \mathbf{I}\,\sigma_e^2\end{bmatrix},$$

where **A** is Wright's numerator relationship matrix, **G** is the (co)variance matrix of the additive genetic random regression coefficients, **I** is an identity matrix whose order is equal to the total number of observations, **P** is the (co)variance matrix of the permanent

environmental random regression coefficients, and σ_e^2 is the variance of random residuals.

Worth mentioning is that in some studies, the random residual variance has been allowed to vary (between observations taken in multiple years, for example). Jamrozik et al., (1997a) modified the residual variance structure $\mathbf{I}\sigma_e^2$ presented above to the following:

$$\text{var}[\mathbf{e}] = \text{diag}\{\sigma_{e_k}^2\}$$

where k is equal to the total number of differing residual variances. In this example, the authors used k = 29 resulting in **e** having 29 different values depending on the number of days in milk which ranged from 1 to 305. Perhaps, another more appropriate method for modeling heterogeneous residual variance is to allow the variance to follow a continuous function (Rekaya et al., 2000). Both methods account for changing residual variance structures, and López et al. (2004) found the two methods to be equivalent. It is important to note that if the assumption of homogeneous residual variance does not hold across all stages of production, a modification should then be made to the model which allows the residual variance to change between those stages of production. Olori et al. (1999) determined the assumption of homogeneity of residual variance will bias the residual variance estimates, leading to over- or under-estimation of heritability values. However, the assumption of homogeneous residual variance has no effect on permanent environmental variance (López-Romero et al., 2003).

Covariance Functions. At approximately the same time the techniques for random regression methodology was being introduced and subsequently implemented, covariance functions were introduced in a series of three papers (Kirkpatrick and Heckman, 1989; Kirkpatrick et al., 1990; Kirkpatrick et al., 1994) with the specific goal of how to account for the changes in the covariance structure between successive observations of longitudinal data. Initial groundwork for the development of the covariance function was first reported by Kirkpatrick and Heckman (1989). They defined the covariance function as the infinite-dimensional counterpart to covariance matrices used in standard multivariate analyses and offered three advantages over the conventional methods. Their three advantages are as follows:

1) Covariance functions have the ability to describe the trait at all points, even if measurements were not taken on specific days, rather than at a finite number of data points;
2) Covariance functions help to reduce errors in calculating the response to selection. Conventional methods only select on a specific age window (for example birth weight or weaning weight), however when selection on a part of the curve is performed, the entire trajectory is changed through the genetic correlation (selecting on increasing birth weight has the correlated effect of increasing weaning weight). Covariance functions help account for the correlated responses observed at other data points as well;
3) Covariance functions estimate parameters more efficiently due simply to the fact that more data points are used in the analysis.

Kirkpatrick et al. (1990, 1994) provided additional insight into the covariance function they introduced in 1989, with examples using a beef cattle growth data set. Calculating the covariance function begins with the standard classical quantitative genetic (co)variance matrix of the traits in question over different time periods, often referred to as **G** (see the multivariate model presented above). Using a beef cattle growth analysis as an example, the genetic (co)variance matrix (**G**) could consist of the additive genetic variance for birth weight and weaning weight. Using this **G**, covariance functions are built by using a smooth curve to interpolate the values of **G** between the measured ages (birth weight and weaning weight). The process starts with the decision as to which smooth curve to use. Kirkpatrick et al. (1990) suggests the use of Legendre polynomials, but states that any orthogonal function could in fact be used. For longitudinal data such as growth, the authors favored polynomials because growth tends to be smooth similar to the curves created using polynomial functions.

A number of sources illustrate the calculation of Legendre polynomial functions. The equations presented here were adapted from Schaeffer (2003). To calculate Legendre polynomials, first we need to define the polynomials:

$$P_0(x) = 1, \text{ and } P_1(x) = x.$$

Then the additional polynomials can be calculated using the recursive formula:

$$P_{n+1}(x) = \frac{1}{n+1}\Big((2n+1)\ x\ P_n(x) - nP_{n-1}(x)\Big).$$

These values are then normalized using:

$$\phi_n(x) = \left(\frac{2n+1}{2}\right)^{0.5} P_n(x).$$

Combining the above equations will result in the following series of Legendre Polynomials:

n = 0	$P_1(x) = x$	$\phi_0(x) = 0.7071$
n = 1	$P_2(x) = \frac{3}{2}x^2 - \frac{1}{2}$	$\phi_1(x) = 1.2247x$
n = 2	$P_3(x) = \frac{5}{2}x^3 - \frac{9}{6}x$	$\phi_2(x) = 2.3717x^2 - 0.7906$
n = 3	$P_4(x) = \frac{35}{8}x^4 - \frac{45}{12}x^2 + \frac{3}{8}$	$\phi_3(x) = 4.6771x^3 - 2.8062x$
n = 4	$P_5(x) = \frac{63}{8}x^5 - \frac{35}{4}x^3 + \frac{15}{8}x$	$\phi_4(x) = 9.2808x^4 - 7.9550x^2 + 0.7955$

This series of normalized polynomials ($\phi_n(x)$) are then put into a matrix **Λ** such that:

$$\mathbf{\Lambda}' = \begin{pmatrix} 0.7071 & 0 & 0 & 0 & 0 \\ 0 & 1.2247 & 0 & 0 & 0 \\ -0.7906 & 0 & 2.3717 & 0 & 0 \\ 0 & -2.8062 & 0 & 4.6771 & 0 \\ 0.7955 & 0 & -7.9550 & 0 & 9.2808 \end{pmatrix}.$$

Legendre polynomials are defined over the interval of -1 to 1 (Kirkpatrick et al., 1990), therefore it is necessary to standardize the ages of the observations to the interval of -1 and 1. The formula used to standardize these ages was presented by Schaeffer (2003) and is defined as follows:

$$t_i^* = -1 + 2\left(\frac{t_i - t_{min}}{t_{max} - t_{min}}\right)$$

where t_i^* is the standardized time, t_i is the time point being standardized, and t_{min} and t_{max} were the minimum and maximum time points or ages represented in the dataset, respectively. Standardized time values are placed in to a matrix **M** such that an example standardized age vector $t_i^* = \begin{bmatrix}-1 & -.25 & .25 & 1\end{bmatrix}^T$ would result in:

$$\mathbf{M} = \begin{bmatrix} 1 & -1 & 1 \\ 1 & -0.25 & 0.0625 \\ 1 & 0.25 & 0.0625 \\ 1 & 1 & 1 \end{bmatrix}$$

for a quadratic polynomial. The first column of the matrix is a column of ones representing the intercept of the curve; the second column is the standardized age while the third column is the standardized age squared for the quadratic term. Fitting higher order polynomials is done by the addition of columns for the additional parameters needed. The next step is to combine the standardized ages and the polynomials into a matrix $\mathbf{\Phi} = \mathbf{M\Lambda}$. Performing this step with the $\mathbf{M}$ defined above and the first three rows (quadratic) of $\mathbf{\Lambda}'$ gives the matrix

$$\mathbf{\Phi} = \begin{bmatrix} 0.7071 & -1.2247 & 1.5811 \\ 0.7071 & -0.30618 & -0.64237 \\ 0.7071 & 0.30618 & -0.64237 \\ 0.7071 & 1.2247 & 1.5811 \end{bmatrix}$$

which when combined with the original genetic (co)variance matrix, using the formula

$$\widehat{\mathbf{C}}_G = \Phi^{-1}\widehat{G}\left[\Phi^T\right]^{-1}$$

results in an estimated coefficient matrix $\widehat{\mathbf{C}}_\mathbf{G}$ from which the covariance function can be formed (Kirkpatrick et al., 1990).

The estimated $\mathbf{C}$ matrix can be used in conjunction with the following covariance function to estimate the covariance between any two measurements taken at any two standardized times denoted t_1 and t_2 (Kirkpatrick and Heckman, 1989; Kirkpatrick et al., 1990; Kirkpatrick et al., 1994):

$$f(a_1,a_2) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}[C_G]_{ij}\,\phi_i\left(t_1^*\right)\phi_j\left(t_2^*\right)$$

where $[\mathbf{C}_G]_{ij}$ is the i^{th} and j^{th} element of the estimated matrix $\widehat{\mathbf{C}}_G$, and $\phi_{i(j)}$ is the Legendre polynomial coefficient for the i^{th} age and j^{th} order. The use of this equation is somewhat limited though given phenotypic measurements are typically measured at n ages. Therefore, only an n × n truncated version of $\mathbf{C}_G$ can be used (Kirkpatrick et al., 1990).

The preceding discussion details the formation of a covariance functions for a full order fit, meaning the number of orthogonal functions estimated (k) equals the number

of ages measured (n) and is equivalent to the multivariate model (Mrode, 2005). Given a situation where a large number of different ages were measured, meaning n becomes large; the problem becomes intractable rather quickly. Kirkpatrick et al. (1990) determined it possible, and in some cases more attractive, to reduce the order of fit (k < n) such that the covariance matrix can be fitted with as few parameters as possible. The reduced order covariance function was found using weighted least squares procedures to identify the simplest orthogonal function in which the reduced (co)variance matrix was not significantly different from the full (co)variance matrix as determined from a χ^2 goodness of fit test. If the reduced (co)variance matrix differed significantly from the full order matrix, the order of the reduced matrix was increased by using higher order Legendre polynomials until the reduced and full matrices did not differ significantly. According to Kirkpatrick et al. (1990), the reduced estimate is the simplest polynomial that is "statistically consistent" with the data. It also smoothes out the fluctuations caused by the sampling error in the initial measurements used to estimate **G**. The authors do caution, however, that this method will exclude higher order terms even if they actually exist if the data is not powerful enough to show their existence.

Random Regression versus Covariance Functions. Meyer and Hill (1997) were the first to show the equivalence of the random regression model to the covariance function, and then Mrode (2005) illustrated this equivalence through the use of an example. He compared the covariance between breeding values calculated from data recorded on an individual animal using both a parametric curve and a set of orthogonal polynomials fitted in a random regression model. The equality of the covariance function to the random regression model allows the estimation of fewer regression coefficients for each source of variation. When used in random regression models, the matrix □ replaces the standard covariate incidence matrix.

Recently, some issues have surfaced concerning random regression models which employ the use of Legendre polynomials as their basis function. The estimated covariance matrices used to calculate genetic variances over the range of data (over the range of lactation for instance) tend to result in genetic variances that are much

higher at the beginning and end of the data range than in the middle (Schaeffer & Jamrozik, 2008). Perhaps this is due to the fact that polynomials place a large amount of emphasis on observations at the extremes, which compounds the problem with higher orders of fit (Meyer, 2005a). Other reported problems with Legendre polynomials being used in random regressions are the poor modeling capabilities of asymmetrical functions, their lack of information to estimate a large number of parameters, and their sensitivity to each of the many different (co)variance parameters (Misztal, 2006).

Splines. Given the problems with the use of polynomials as a basis function in random regression models discussed by Misztal (2006) and Schaeffer & Jamrozik (2008), several different alternatives such as fractional polynomials (Robert-Granié et al., 2002), cubic smoothing splines (White et al., 1999), and B-splines (Torres & Quaas, 2001; Meyer, 2005b) have been proposed. Spline functions are defined as piecewise polynomials which join together at "knots" and are continuous across the range of data (Wold, 1974). As a result, they do not suffer from the same problems as polynomials where their behavior in one small area determines their behavior across the entire range of data. Since splines are defined as "piecewise polynomials" they represent smooth curves between each knot.

Ruppert et al. (2003) describes simple spline basis functions as an extension of the following standard simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

where y_i is the observed value of the i^{th} trial, x_i is the predictor variable of the i^{th} trial, β_0 and β_1 are regression coefficients corresponding to the y intercept and slope of the regression line, respectively, and e_i is the random error term with mean equal to 0 and variance equal to σ_e^2. This model can be easily extended to higher order polynomials through the addition of one more regression coefficient and predictor variable squared such that:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i.$$

The quadratic simple linear regression model presented above would result in an **X** incidence matrix for fitting the regression of:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix}.$$

Modification of these models for the inclusion of "knots" or points where the piecewise polynomials join together is a rather simple task accomplished by the addition of K columns of $(x_i - \kappa)_+$ where κ is specific knot and "+" refers to the positive section of the function, meaning negatives values of $(x_i - \kappa)$ are included as zero. These values are included in the general simple linear regression in such a manner where:

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} (x - \kappa_k)_+ ,$$

and in the quadratic version of this model as:

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \sum_{k=1}^{K} b_k (x - \kappa_k)_+^2.$$

The **X** incidence matrix associated with the above quadratic spline equation would then be modified to be:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & (x_1-\kappa_1)_+^2 & \cdots & (x_1-\kappa_K)_+^2 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & (x_n-\kappa_1)_+^2 & \cdots & (x_n-\kappa_K)_+^2 \end{bmatrix}.$$

These "modified" **X** matrices are then included in the Least Squares normal equations as a substitute for the standard simple linear regression **X** incidence matrices. As a result, standard Least Squares regression statistical properties apply and fitted values can be found by solving the normal equations:

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

The spline basis functions presented above are referred to as truncated power bases of degree p, and useful for understanding the mechanics of spline based regression. They can be used in practice if knots are carefully chosen or a penalized fit (inclusion of a roughness penalty or a value which penalizes fits which are too rough, resulting in a smoother result) is used (Ruppert et al., 2003). Truncated power base functions are at a disadvantage when it comes to orthogonality, meaning numerical instability can result if too many knots are used and the roughness penalty is too small. It has been suggested that the use of equivalent bases such as B-splines or natural

cubic smoothing splines with more stable numerical properties would be desirable (Eilers and Marx, 1996; Ruppert et al., 2003).

Given the piecewise nature of spline bases, some of the problems associated with random regression using Legendre polynomials such as instability at the extremes, seem to be avoided. In 1999, Hill and Brotherstone reported that splines can be included rather easily in the standard mixed model framework and when compared to random regression models, they include more random effects but require fewer (co)variance parameters. Splines also have the advantage of quicker convergence over Legendre polynomials, which may be due to the fact that spline coefficients are more sparse than their polynomial counterparts (Misztal, 2006). One of the most important questions when using spline bases seems to be related to the number of knots needed to accurately describe the data as well as where to place these knots. The use of too many knots will increase the complexity of the model, while the use of too few will reduce accuracy. Misztal (2006) suggests the following guidelines for choosing proper knot placement:

1) Choose knots in such a manner that they encompass the extremes observed in the data.
2) Choose knots in a way that the correlations between knots is in the range of 0.6 to 0.8.

These two suggestions will result in knots being placed close together around areas that have the largest data density (i.e. birth weight, weaning weight, etc.), and will also result in a larger concentration of knots in areas where the data is changing more rapidly.

Until very recently, use of spline based regression techniques by quantitative geneticists in the livestock industry had been almost non-existent. Spline basis functions have been used in the analysis of a number of traits, and as with the random regression and covariance function models, they were first proposed for the analysis of dairy cattle test day records. They have been incorporated into fixed regressions to model the lactation curve in the analysis of dairy cattle test day records (Druet et al., 2003), as well as the modeling of curves for estimated breeding values (White et al., 1999).

The use of splines for the analysis of beef cattle data seems, so far, limited to the analysis of growth traits. Meyer (2005b) used quadratic B-splines to analyze Angus growth data from birth to 820 days of age with knots at 0, 200, 400, 600 and 821 days of age. She found that the B-splines lend themselves well to the modeling of growth data, but they tend to be susceptible to irregularities in the distribution and sparseness of the data. Using simulated beef cattle growth data, Bohmanova et al. (2005) found that despite the fact that splines are simpler with fewer parameters than Legendre polynomials, they are just as accurate (within 0.2%). A series of studies was conducted in 2005 and 2006 investigating the feasibility of using spline basis functions in random regression models with the application to large scale genetic evaluations (Iwaisaki et al., 2005; Robbins et al., 2005; Bertrand et al., 2006). In this set of studies, it was determined that random regression using spline bases is a feasible alternative to random regression with Legendre polynomial bases as well as the more contemporary multivariate model.

Conclusions and Implications to the Beef Industry. A topic of much discussion today among beef cattle geneticists is how to add accuracy to existing genetic evaluations. While much of this discussion is focused on the incorporation of DNA marker information into current genetic evaluation schemes (Schaeffer, 2006; Goddard and Hayes, 2007; Kachman, 2008), increasing the number of useable records on a sire is another possibility for adding accuracy to a sire's EPD. Current evaluation methodology for growth uses a multivariate mixed model which treats weights recorded at successive ages as separate but genetically correlated traits. If the number ages at which observations are measured is high, these problems can become very large and intractable rather quickly. As a result, the Beef Improvement Federation (BIF) recommends that weights be standardized to a specified age (for example, weaning weights are typically adjusted to 205d) or weights fall into specified age ranges (BIF recommended age range for weaning weight is 160 to 250 days) in order to be included in a genetic evaluation (BIF, 2002). Often management decisions, such as early weaning strategies, lead to weights being recorded outside these recommended age ranges, and therefore render them unusable for genetic evaluation purposes. The use

of random regression techniques has the ability to incorporate observations taken at any number of ages. This is an advantage over conventional genetic evaluation methods due to the fact that increasing the amount of useable data has the effect of increasing the accuracy of the genetic prediction which can lead to an increased rate of genetic change (Williams et al., 2009).

Random regression methodology also has the potential to have a larger impact on the beef industry than just in the analysis of growth traits. It has the ability to identify cattle that require fewer days to reach their finish endpoint in the feedlot, a trait (group of traits) which has long been identified as being economically relevant (Lindholm and Stonaker, 1957; Golden et al., 2000). Besides the ability of random regression models to use data measured at a number of ages, the resulting EBV obtained from these models are EBV for regression curves. Given this, EBV could be calculated for any age or any number of days on feed. Kuehn (2000) presented an equation for the calculation of any customized EBV where individual animals were estimated using a linear regression as is shown below:

$$\text{EBV(age or weight)} = b_0 + b_1*(\text{desired age})$$

where b_0 is the EBV for the intercept and b_1 is the linear EBV for each individual sire. Following Kuehn, Julieu (2003) showed sufficient genetic variation existed for a selection tool expressed as days to finish weight. The EBV obtained from a random regression model, on the surface, have the potential to cause confusion, especially if higher order polynomials are used. Therefore, information packaging and decision support of genetic evaluations which incorporate random regression approaches should be carefully considered.

Random regression techniques have the potential to greatly influence beef cattle genetic evaluation techniques. Their ability to incorporate recorded data from any number of ages into a single evaluation, by properly accounting for the changing covariance structure between observations, has the potential to have a large impact on genetic evaluation methodology with little to no change on current data recording schemes. As such, increases in the accuracy of an evaluation will be seen as these techniques become more widely used in national beef cattle evaluations.

Literature Cited

Beef Improvement Federation. 2002. BIF Guidelines for Uniform Beef Improvement Programs. (8th Ed.). Athens, GA.

Bertrand, J. K., I. Misztal, K. R. Robbins, J. Bohmanova, and S. Tsuruta. 2006. Implementation of random regression models for large scale evaluations for growth in beef cattle. In: Proc. 8th World Congr. Appl. Livest. Prod. Belo Horizonte, Brazil.

Bohmanova, J., I. Misztal, and J. K. Bertrand. 2005. Studies on multiple trait and random regression models for genetic evaluation of beef cattle for growth. J. Anim. Sci. 83:62-67.

Carvalheira, J. G. V., R. W. Blake, E. J. Pollak, R. L. Quaas, and C. V. Duran-Castro. 1998. Application of an autoregressive process to estimate genetic parameters and breeding values for daily milk yield in a tropical herd of Lucerna cattle in United States Holstein herds. J. Dairy Sci. 81:2738-2751.

Druet, T., F. Jaffrézic, D. Boichard, and V. Ducrocq. 2003. Modeling lactation curves and estimation of genetic parameters for first lactation test day records of French Holstein cows. J. Dairy. Sci. 86:2480-2490.

Eilers, P. H. C., and B. D. Marx. 1996. Flexible smoothing with B-splines and penalties. Statistical Science. 11:89-121.

Foster, J. J., E. Barkus, and C. Yavorsky. 2006. Understanding and Using Advanced Statistics: A practical guide for students. 1st ed. Sage Publications Ltd.

Golden, B. L., D. J. Garrick, S. Newman, and R. M. Enns. 2000. Economically relevant traits, a framework for the next generation of EPDs. Proc. Beef Improvement Federation, Wichita, KS, pp. 2-13.

Goddard, M. E. and B. J. Hayes. 2007. Review Article: Genomic Selection. J. Anim. Breed. Genet. 124:323-300.

Harville, D. A. 1979. Recursive estimation using mixed linear model with autoregressive random effects. In: Variance Components and Animal Breeding. Proc. Conf. In honor of C. R. Henderson. Pp. 157-179. Cornell Univ., Ithaca, NY.

Henderson, C. R. and R. L. Quaas. 1976. Multiple trait evaluation using relatives' records. J. Anim. Sci. 43:1188-1197.

Henderson, C. R. 1982. Analysis of covariance in the mixed model: higher-level nonhomogeneous and random regressions. Biometrics. 38:623-640.

Hill, W. G. and S. Brotherstone. 1999. Advances in methodology for utilizing sequential records. British Soc. Anim. Sci. Occaisonal Pub. 24:55-61.

Interbull. 2000. National Genetic Evaluation Programmes for Dairy Production Traits Practised in Interbull Member Countries 1999-2000. Department of Animal Breeding and Genetics, Uppsala, Sweden, Bulletin 24.

Iwaisaki, H., S. Tsuruta, I. Misztal, and J. K. Bertrand. 2005. Genetic parameters estimated with multitrait and linear spline-random regression models using Gelbvieh early growth data. J. Anim. Sci. 83:757-763.

Jamrozik, J., G. J. Kistemaker, J. C. M. Dekkers, and L. R. Schaeffer. 1997a. Comparison of possible covariates for use in a random regression model for analyses of test day yields. J. Dairy. Sci. 80:2550-2556.

Jamrozik, J., L. R. Schaeffer, and J. C. M. Dekkers. 1997b. Genetic evaluation of dairy cattle using test day yields and random regression model. J. Dairy Sci. 80:1217-1226.

Jennrich, R. I., and M. D. Schluchter. 1986. Unbalanced repeated-measures models with structured covariance matrices. Biometrics. 42:805-820.

Jublieu, J. S. 2003. The use of random regression models to predict days to finish in beef cattle. M. S. Thesis. Colorado State University, Fort Collins, CO.

Kachman, S. D., and R. W. Everett. 1993. A multiplicative mixed model when the variances are heterogeneous. J. Dairy. Sci. 76:859-867.

Kachman, S. D. 2008. Incorporation of marker scores into national genetic evaluations. Proc. Genetic Prediction Workshop. Kansas City, MO.

Kirkpatrick, M. and N. Heckman. 1989. A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. J. Math Biol. 27:429-450.

Kirkpatrick. M., D. Lofsvold, and M. Bulmer. 1990. Analysis of the inheritance, selection and evolution of growth trajectories. Genetics. 124:979-993.

Kirkpatrick, M., W. G. Hill, and R. Thompson. 1994. Estimating the covariance structure during growth and ageing, illustrated with lactation in dairy cattle. Genet. Res. Camb. 64:57-69.

Koots, K. P., J. P. Gibson and J. W. Wilton. 1994. Analyses of published genetic parameter estimates for beef production traits. 2. Phenotypic and genetic correlations. Anim. Breed. Abstr. 62:825-853.

Kuehn, L. A. 2000. Parameterization of random regression models for beef cattle data sets. M. S. Thesis. Colorado State University, Fort Collins, CO.

Laird, N. M., and J. H. Ware. 1982. Random-effects models for longitudinal data. Biometrics. 38:963-974.

Lindholm, H. B. and H. H. Stonaker. 1957. Economic importance of traits and selection indexes for beef cattle. J. Anim. Sci. 16:998-1006.

López-Romero, P., R. Rekaya, and M. J. Carabaño. 2003. Assessment of homogeneity vs. heterogeneity of residual variance in random regression test-day models in a Bayesian analysis. J. Dairy. Sci. 86:3374-3385.

López-Romero, P., R. Rekaya, and M. J. Carabaño. 2004. Bayesian comparison of test-day models under different assumptions of heterogeneity for the residual variance: the change point technique versus arbitrary intervals. J. Anim. Breed. Genet. 121:14-25.

Meyer, K. 2004. Scope for a random regression model in genetic evaluation of beef cattle for growth. Livest. Prod. Sci. 89:69-83.

Meyer, K. 2005a. Advances in methodology for random regression analyses. Aust. J. Exp. Agric. 45:847-858.

Meyer, K. 2005b. Random regression analyses using B-splines to model growth of Australian Angus cattle. Genet. Sel. Evol. 37:473-500.

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.

Meyer, K. and M. Kirkpatrick. 2005. Up hill, down dale: quantitative genetics of curvaceous traits. Phil. Trans. R. Soc. B. 360:1443-1455.

Misztal, I. 2006. Properties of random regression models using linear splines. J. Anim. Breed. Genet. 123:74-80.

Mrode, R. A. 2005. Linear models for the prediction of animal breeding values. 2nd ed. CABI Publishing Company. Cambridge, MA.

Olori, V. E., W. G. Hill, and S. Brotherstone. 1999. The structure of the residual error variance of test day milk yield in random regression models. Workshop on Computational Breeding, Finland. March, 1999. Pp. 18-20.

Pander, B. L., W. G. Hill, and R. Thompson. 1992. Genetic parameters of test day records of British Holstein-Friesian heifers. Anim. Prod. 55:11-21.

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

Rekaya, R., M. J. Carabaño, and M. A. Toro. 2000. Assessment of heterogeneity of residual variances using change point techniques. Genet. Sel. Evol. 32:383-394.

Robbins, K. R., I. Misztal, and J. K. Bertrand. 2005. A practical longitudinal model for evaluating growth in Gelbvieh cattle. J. Anim. Sci. 83:29-33.

Robert-Granié C., E. Maza, R. Rupp, and J. L. Foulley. (2002). Use of fractional polynomial for modeling somatic cell scores in dairy cattle. In: Proc. 7th World Congr. Appl. Livest. Prod. CD-ROM, comm.. no 16-05, Montpellier, France.

Ruppert, D., M. P. Wand, and R. J. Carroll. 2003. Cambridge series in statistical and probabilistic mathematics: Semiparametric Regression. Cambridge University Press. New York, NY.

Schaeffer, L. R. 2003. Random regression models. ANSC637 Course Notes – Quantitative genetics and animal models. Available at: http://www.aps.uoguelph.ca/~lrs/ANSC637/LRS14/LRS14.pdf. Accessed Jan. 8, 2009.

Schaeffer, L. R. 2004. Application of random regression models in animal breeding. Livest. Prod. Sci. 86:35-45.

Schaeffer, L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123:218-223.

Schaeffer, L. R. and J. C. M. Dekkers. 1994. Random regressions in animal models for test-day production in dairy cattle. In: Proc. 5th World Congr. Appl. Livest. Prod. XVIII. Pp. 443-446.

Schaeffer, L. R. and J. Jamrozik. 1996. Multiple-trait prediction of lactation yields for dairy cows. J. Dairy Sci., 79:2044-2055.

Schaeffer, L. R. and J. Jamrozik. 2008. Random regression models: a longitudinal perspective. J. Anim. Breed. Genet. 125:145-146.

Torres, R. A., and R. L. Quaas. 2001. Determination of covariance functions for lactation traits on dairy cattle using random-coefficient regressions on B-splines. J. Anim. Sci., 79(Suppl. 1):112 (Abstr.)

White, I. M., R. Thompson, and S. Brotherstone. 1999. Genetic and environmental smoothing of lactation curves with cubic splines. J. Dairy. Sci. 82:632-638.

Wiggans, G. R. and M. E. Goddard. 1996. A computationally feasible test day model with separate first and later lactation genetic effects. Proc. New Zealand Soc. Anim. Prod. 56:19-21.

Wiggans, G. R. and M. E. Goddard. 1997. A computationally feasible test day model for genetic evaluation of yield traits in the United States. J. Dairy. Sci. 80:1795-1800.

Williams, J. L., D. J. Garrick, and S. E. Speidel. 2009. Reducing bias in maintenance energy EPD by accounting for selection on weaning and yearling weights. J. Anim. Sci. (In Press).

Wold, S. 1974. Spline functions in data analysis. Technometrics. 16:1-11.

Wright, S. (1922). Coefficients of inbreeding and relationship. American Naturalist. 56:330-338.