A PRACTICAL LONGITUDINAL MODEL FOR EVALUATION OF GROWTH TRAITS

by

KELLY ROBBINS

(Under the Direction of Joseph Keith Bertrand and Ignacy Misztal)

ABSTRACT

For the first study a multiple-trait model (MTM), a random regression model utilizing Legendre polynomials (RRML), and a random regression model construct with linear spline functions (RRMS) were applied for analysis of national beef cattle growth data. The impact of the additional information included in the RRML and RRMS were examined through correlations of random effect predictions. Results showed decreases in correlations between MTM and both RRM when additional information was incorporated into RRM analysis.

The second study focused on the modeling of fixed effects within the frame work of a RRM. Models utilizing polynomials and two-dimensional splines were evaluated via cross validation based on the average squared error (ASE), R^2, and percent bias. Due to the nonlinearity of two-dimensional splines, weighted spline extrapolation had to be used outside the two-dimensional grid. Results showed comparable performance between polynomials and two-dimensional splines.

INDEX WORDS: Multiple-trait model, Random regression, Legendre polynomials, Splines, Cross validation, Two-dimensional splines

A PRACTICAL LONGITUDINAL MODEL FOR EVALUATION OF GROWTH TRAITS

By

KELLY ROBBINS

B.S., The University of Tennessee, 2002

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment
of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2005

A PRACTICAL LONGITUDINAL MODEL FOR EVALUATION OF GROWTH TRAITS

by

KELLY ROBBINS

Co-Major Professors: Joseph Keith Bertrand
Ignacy Misztal

Committee: Romdhane Rekaya
Samuel Aggrey

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2005

ACKNOWLEDGEMENTS

I would like to thank Dr. Bertrand and Dr. Misztal for their guidance and assistance. Without their input and knowledge I could not have finished. I would also like to thank Dr. Rekaya for his valuable input and motivation. Finally, I would like to thank the faculty, staff, and graduate students in the Animal and Diary Science Department at The University of Georgia for creating an enjoyable working environment.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Multiple-trait models (MTM) currently used to evaluate growth in beef cattle treat all records for a trait as being measured at the same age and having the same parameters. This creates the need for age adjustments, establishment of age ranges, and the subsequent elimination of records measured outside these ranges. These characteristics of MTM lead to the loss of records, particularly for early weaned animals. Random regression models (RRM) utilize continuous (co)variance functions to model records taken at any age, thus eliminating the need for MTM adjustments.

RRM utilizing orthogonal Legendre polynomials (RRML) have been widely used for the modeling of random effects. Using methods that derive RRML parameters directly from MTM (co)variance matrices, stable and reliable functions can be obtained. While much of the focus on longitudinal models has centered on RRML, linear splines present a simple alternative. Splines are computationally simple and have the ability to use MTM parameters directly. A spline fits a series of functions through control points referred to as knots. The use of linear spline functions with few knots can eliminate the numerical problems associated with RRML and decrease the cost of implementation.

The accurate modeling of fixed effects in any analysis is crucial for eliminating bias in random effect predictions. When evaluating growth, two continuous covariates are commonly modeled, the age of dam effect (AOD) and the animal's age. Currently records are collected for MTM and are clustered around birth, weaning (205 days), and yearling (365 days) ages. With such age distributions, nested polynomial functions seem to be an obvious choice for the modeling of these effects. However, as RRM models become more widely accepted, larger and more continuous age ranges will be recorded for evaluation. This could make nesting more difficult and outlying age records could make polynomials vulnerable to artifacts.

An alternative to polynomial regressions are splines. The piecewise nature of splines makes them more resistant to artifacts of data. While one-dimensional splines provide a more robust model, they still require nesting when modeling AOD and age of animal. Two-dimensional splines can provide a generalized and robust model for fixed effects. The selected two-dimensional knots provide automatic nesting and implicit modeling of interactions while reducing the effects of outlying records.

The objective of these studies was to develop a robust and practical longitudinal model for the evaluation of growth in beef cattle through effective modeling of both fixed and random effects.

CHAPTER 2

REVIEW OF LITERATURE

The use of random regression models (RRM) to model longitudinal data has been discussed since the early 1980's when introduced by Henderson (1982). Initial interest in the field of breeding genetics was focused mainly around the evaluation of test day milk yield in dairy cattle (Schaeffer and Dekkers**,** 1994), but RRM are suited for modeling of any type of longitudinal data. This has led to applications of RRM to growth data in animal production. Anderson and Pederson (1996) examined the use of RRM to growth and feed intake curves in pigs. Meyer and Hill (1997) discussed the use of RRM to model growth in beef cattle data. Given the large scope for RRM, this discussion will mainly focus on its use in modeling random effects in the context of breeding and genetics, and expand to multiple applications for discussion of fixed effects.

RRM utilize (co)variance functions to model traits that have measurements that change over time. These functions model the changes in a trait, as well as the correlation between measurements taken at different time points. RRM can be modeled with any continuous function such as polynomials, orthogonal polynomials, and splines. In an early application of RRM**,** Kirpatrick et al. (1990) examined the modeling of additive genetic (co)variance functions using full and reduced models. RRM utilizing orthogonal Legendre polynomials (RRML) were chosen to model the function because, unlike spline functions, coefficients from the smooth orthogonal polynomials provide information on the changes in genetic variation within growth trajectories. Using techniques of stacking matrices, parameters for RRML were derived from multivariate (co)variance matrices. The use of (co)variance functions gave two distinct advantages over multivariate analysis; the first being the fact that RRM incorporates spacing in terms of age between weight records, and secondly, with RRM the ages at which selection pressure should be applied do not need to coincide with the ages at which records are recorded.

Legendre polynomials partition each order of the regression into orthogonal components. This yields coefficients with reduced correlations. This desirable property makes it an appealing choice for use in RRM. As a result, the majority of RRM modeling has been done with orthogonal polynomials. Van Der Werf et al. (1998) examined the use of Legendre polynomials to model (co)variance functions to evaluate yield traits for milk production in dairy cattle. Data were collected from 30 randomly selected Australian herds. The data set contained 13,109 records on 1,903 Friesian cows. Parameters for the covariance function were estimated by transformation of parameters for predetermined periods of lactation and directly from the data using RRML. The first method gave more stable estimates, especially at the extremes of the lactation curve. Goodness of fit procedures showed that RRML and RRM based on lactation curves provided very similar fits to the data. Additionally the equivalence of RRM and (co)variance functions was shown.

Tijani et al. (1999) computed parameter estimates for covariance functions on test day yield. A RRML was fit to first lactation records from 17,190 cows from 37 herds in Wisconsin and Pennsylvania. Parameters for the RRML were derived from the (co)variance matrices for a four stage model in which lactation was divided into 75d intervals. Using the (co)variance function, twelve 25d lactation stages were created from the four lactation stage model. Estimates of genetic and environmental parameters demonstrated the ability of RRML to interpolate and extend (co)variances across the entire lactation. This property makes the RRM an ideal model for the evaluation of test day records.

A comparison of four different test day models for milk yield traits was conducted by Lindauer et al. (2003). Milk, protein, and, and fat yield traits from 1,049248 Finnish dairy cattle were evaluated with four different models. Full and reduced rank RRM, a repeatability model, and a multiple trait (MTM) lactation yield model were each implemented. The mean squared error (MSE), correlations of breeding values, and sire rankings were used to evaluate each model. MSE was lowest for the full rank RRM with very similar MSE for the reduced rank model. Correlations of breeding values were highest for the full and reduced rank RRM. Correlations between RRM and the MTM lactation yield model were > .98 for active bulls. With young animals, correlations dropped and significant re-rankings were observed. The high correlations and low re-rankings between the full and reduced rank RRM show that rank reduction has little effect on model predictions. This combined with the improved convergence and decreased memory requirements of the reduced rank RRM, make it a practical choice for modeling test day records.

In addition to modeling test day milk yield traits, the use of RRML for modeling growth has also been the subject of intensive research in recent years**.** Anderson and Peterson (1996) constructed a polynomial regression with RRM coefficients to model growth and feed intake curves in pigs. The data consisted of 96 castrated males and 96 guilts from a single herd. When pigs reached 30 kg they were weighed twice a week until they reached 95 or 115 kg. Feed intake was recorded for individual pigs by feeding machines. Though plots of weight records showed linear growth through the test period, it was found that forth order polynomial models provided the best fit. Additionally it was found that growth rate was not constant, with maximal rates of growth attained between 30 and 40 days of age.

The use of RRM to model growth in beef cattle has garnered particular interest in the field of breeding and genetics. Alberquerque and Meyer (2001) explored the order of fit required for RRM analysis of beef cattle growth data. RRML of varying orders were applied to Brazilian Zebu cattle records. The data set contained only animals with three weights. The best models, as identified by Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC), had forth and sixth order regressions on animal direct effects, respectively. The results of the model selection strongly implied that adequate modeling of random effects for growth in beef cattle required the use of fourth order or higher polynomial functions.

Currently growth in beef cattle is modeled by multivariate models referred to as multiple trait models (MTM), but the longitudinal nature of the data makes the use of RRM a more theoretically sound model for evaluation of weight data (Kirkpatrick et al., 1990). Meyer and Hill (1997) compared the properties of RRM and MTM for evaluation of growth in beef cattle. It was noted that an MTM with the number of traits equal to the number of ages would be greatly over-parameterized, while RRM could use a function model (co)variances at any age without over-parameterization. Furthermore, eigenvalues from a covariance matrix could provide information on how mean growth curves will change under selection pressure. It was also noted that (co)variance function parameters could be easily estimated from the data using maximum likelihood methodology.

Meyer (2004) used simulated data to compare the accuracies of RRML and MTM in the prediction of breeding values (BV) for growth. Data were simulated assuming a cubic function on age for random effects. Accuracies of the MTM and RRML were calculated as correlations between the true and simulated values. It was found that when analyzing the same records, the RRML was consistently more accurate than the MTM. This was attributed to the more correct modeling of the (co)variances in the RRML. When incorporating data outside the MTM ranges the increases in the accuracies of the RRML, when compared to RRML utilizing only MTM data, depended on the number of records located beyond MTM ranges. Based on these findings it was concluded that when large numbers of records were located outside the MTM ranges, the RRML's ability to incorporate such records could yield more accurate BV predictions.

While the RRML provides more appropriate modeling of (co)variance, applications of RRML to growth, as with milk yield (Van Der Werf, 1998), have revealed potential problems concerning the estimation of parameters for Legendre polynomials. Nobre et al. (2003a) estimated genetic parameters for Nelore cattle using REML under the framework of RRML and MTM. Two data sets were used for parameter estimation. One data set contained only animals with complete records collected and the other contained animals with missing records. Parameter estimates from data containing missing records were compared to estimates from complete data to determine the effects of missing information on RRML parameter estimates. It was found that estimates from RRML using incomplete data show large increases in estimates after 600d of age. It was concluded that estimates of parameters from RRML in sparse areas of data were subject to artifacts of the data.

Nobre et al. (2003b) applied both MTM and RRML to growth records of 619,989 Nelore cattle. The parameters for both models were from Nobre et al. (2003a). In order to achieve better numerical properties for the RRML, parameter matrices were diagonalized and order was reduced based on eigenvalues that were close to zero. When diagonalization was not performed correlations between iterative solvers and programs that solved via inversion were very low.

With diagonaliztion and strict convergence criteria, these correlations jumped to .998. Correlations of random effects between MTM and RRML were lower than expected, and were likely a result of the differences in parameter estimates between the two models.

In the absence of comparable parameter estimates it is difficult to compare the performance of MTM and RRM when applied to real data. Legarra et al. (2004) notes that unlike milk yield records, beef cattle records tend to be clustered around MTM ages. This clustering of data makes RRML particularly vulnerable to artifacts when estimating parameters. To eliminate this problem a method to derive more accurate and practical parameters was examined. Using techniques similar to those used by Kirkpatrick et al. (1990; 1994), RRML parameters were derived directly from MTM parameters. The residuals were split into permanent and temporary environmental components. Deriving the parameters rather than estimating them from the data eliminated problems arising from artifacts, as well as assuring that RRML and MTM parameters were equivalent, which allowed more informative comparisons of model performances with real data.

While much attention has been focused on the use of RRML, splines present a simple alternative to model growth data. Splines are simple piecewise functions that are fit through control points referred to as knots. The ability of splines to use available MTM parameters eliminates problems associated with parameter estimation by RRML. Wold (1974) explored the advantages of using spline functions, arising from their piecewise nature. It was acknowledged that the localizing effect of the piecewise functions makes splines superior for describing disjointed data. Additionally, it was stated that the use of simple polynomials to construct splines makes them computationally simple. Due to the splines' flexibility and local properties, Wold concluded that splines were excellent tools for analysis of continuous data.

A study conducted by Huisman et al. (2002) examined the use of splines to fit (co)variance functions. Both RRML and RRM utilizing a spline function (RRMS) were applied to the weight data of pigs. Sire models were used to evaluate 1,315 boars with weights taken at three periods. AIC and BIC were used to evaluate the performance of each model. It was found that both the RRML and RRMS performed better than MTM. RRMS required fewer parameters to model the data than did the RRML, and was favored by AIC. The results of this study showed that RRM provided the best fit to pig data. Additionally, results indicated that RRMS was a viable alternative to RRML.

Bohmanova et al. (2005) conducted a simulation study to performance of MTM, RRML, and RRMS using linear splines with growth data. Three non-overlapping generations of animals were simulated with random mating of males and females. Random effects were simulated with cubic Legendre polynomials. Four data sets were created containing three records per animal measured at exact intervals, three measurements over varying intervals, five measurements at exact intervals, or five records at varying intervals. Comparisons of accuracies showed that both RRML and RRMS performed better than MTM when ages were measured over uneven intervals. The performance of RRMS was comparable to that of RRML, though RRML had slightly higher accuracies with data containing five records measured at uneven intervals. Plots of (co)variance functions showed that RRML yielded smoother functions, but addition of knots to the spline could reduce such differences.

Much of the work on RRM has focused on the modeling of random effects. However, correct modeling of fixed effects is of equal importance. Schaeffer (2004) stresses the importance of fixed effect modeling in the context of RRM and suggest that careful attention be paid to fixed effect functions. Modeling of fixed effects in RRM has been done largely through the use of polynomial regressions on age. Jamrozik et al. (1997) used a model with regressions on both fixed and random effects to analyze Holstein milk yield traits. There were 5.1 million test day milk yield records collected from four regions in Canada. A subset of the original data set containing only first lactation records was used for evaluation. A cross classified herd test day effect and polynomials nested in 32 subclasses were used to model fixed effects. While polynomials of the same form were used to model both fixed and random effects, it noted that covariables for fixed and random effects could differ. Further research in this area was recommended.

Strabel and Misztal (1999) used single- and two-trait RRML to estimate parameters for milk yield traits in Polish Black and White cattle. Data contained 131,985 first and second parity records on 10,746 cows. Fixed effect models included a cross classified herd test day effect and a polynomial regression with three covariates. The polynomial regression was nested in age-season classes and was intended to model the mean lactation curve within each class.

In an application of RRML to modeling of growth in beef cattle, Meyer (2001) analyzed 42,860 records on 7,185 animals from two herds collected from a selection experiment. RRML with orders of fit up to six were examined. To insure direct comparisons of models, all RRML used the same fixed effect equations. The fixed effect model included a cubic orthogonal polynomial regression on age. Based on previous investigations into fixed effects, it was found that the use of higher order polynomial regressions on age yielded negligible decreases in residual sums of squares.

As with random effects, splines can provide an alternative to polynomial regressions when modeling fixed effects. Guo (2002) showed that splines can effectively model both fixed and random effects using mixed model equations. Unlike polynomials, in which a small subset of data can affect an entire function, splines are defined piecewise, thus localizing the effects of sparse data (Wold, 1974). An additional advantage of splines, resulting from their piecewise construction, is the ability to use computationally simple polynomials to model complex functions (Molinari et al., 2002). Splines have been examined in the context of fixed effect modeling in animal breeding and genetics and modeling other types of biological data.

Druet et al. (2003) examined several functions for the modeling fixed lactation curves of French Holstein cows. Ten data sets with a total of 776,858 test day records on 95,823 cows were used in analysis. Fifth order Legendre polynomials, the Wilkmink curve, the Ali-Schafaeffer curve, a fixed classes curve, and regression splines were chosen to model fixed lactation curves. The five models were compared based on mean sum of squares of the residual, the mean residual, log likelihoods, BIC, and AIC. The regression splines and the fixed classes had mean residuals at or very near zero throughout the lactation while all other curves showed fluctuating bias. Regression splines showed little local variation as opposed to the fixed classes model. It was concluded that regression splines were a compromise between the fixed classes, non-linear functions, and the polynomial regression. The regression splines required limited parameters, had good flexibility, were smooth, and had limited sensitivity to the data. These properties led to the conclusion that regression splines provided the best modeling of fixed lactation curves.

In a study of HIV dynamics, Wu and Zhang (2002) explored the use of natural cubic splines in modeling HIV infection over the course of several years. The performance of cubic splines models for short term and long term HIV dynamics were compared to parametric models traditionally used to model short-term dynamics. Data was collected from 48 HIV-infected patients undergoing highly active antiretroviral therapy. Viral load was measured on days 0, 2, 7, 10, 14, 21, and 28 as well as weeks 8, 12, 24, and 48 after initial treatments. Both AIC and BIC favored the cubic spline model containing one internal knot. The models provided comparable short-term estimates, but only the spline model was capable of modeling long-term effects. Based on these results it was concluded that the spline model significantly outperformed the parametric biexponential model.

Rosenburge et al. (2003) analyzed risks associated with alcohol consumption and oral cancer amongst African Americans in a data set containing 194 cases and 203 controls. Several smoothing and linear spline models were used to analyze the data and compared using AIC. It was found that for both the smoothing and linear splines, two segment functions with an internal knot placed at the median of the data performed the best. Though the smoothing splines provided fluent curves, the AIC for both the linear and smoothing splines were very similar. It is concluded that the use of splines to model risk factors was a practical advancement over models using traditional regression techniques.

Molinari et al. (2002) examined the use of splines to analyze oxidation of low density lipoproteins. Data was collected by continuously recording oxidation of low density lipoproteins over a period of 400 minutes. Data was then analysed using a linear spline with three knots. The spline model was fit using traditional least square equations. It was found that the use of splines provided a computationally simple model for estimation of parameters associated with oxidation of low density lipoproteins.

In a study on the use of splines to model pharmacodynamic data, Aarons et al. (2004) fitted a longitudinal spline model to data from a double-blind, placebo-controlled study. The data was collected from 100 hundred hypertension patients ranging 21 to 79 years of age receiving one of four CR-verapamil doses. The longitudinal spline was constructed using a cubic spline with 11 knots. To test the ability of the spline to handle sparse data, the original data set was reduced by 66%. Comparisons of parameter estimates from the full and reduced data set showed little change in these estimates. The results demonstrated the longitudinal spline's ability to handle sparse data effectively.

The use of two-dimensional splines provides an expansion to one-dimensional splines along a second plane. Two-dimensional splines allow the joint modeling of variables that have dependent curves. The most common use of two-dimensional splines has been in the form of thin plate splines used for engineering and graphical applications (Meinguet, 1979). The modeling of continuous functions along the axes of two variables can provide considerable advantages over one-dimensional splines which model along the axis of a single variable, thus requiring no nesting and providing implicit modeling of interactions when variables are not independent.

Bookstein (1989) describes the thin plate spline as a two-dimensional generalization of the one-dimension cubic spline that models the bending of a thin plane of space between sets of landmark points or knots. When used for biological applications, it is noted that, these knots can be thought of as combinations of two variables that have some meaningful interpretation. It is shown that the two dimensional spline can be broken down to a linear part, as well as, a nonlinear component that provides a smoothing effect between knots. In an application to x-ray data from patients suffering from Apert Syndrome, the mean deviations of eight landmarks in the upper jaw were modeled with the thin plate spline. The x and y variables represented the location landmark points in the jaw of patients suffering from the syndrome. The use of the two-dimensional spline allowed easy modeling of the mean deformations in the upper jaw resulting from Apert Syndrome.

It has been shown that RRM can provide theoretically sound models for analysis of longitudinal data. Due to artifacts of data, models that derive parameters from MTM type (co)variance matrices, rather than from the data directly, provide the most reliable parameter estimates. Both splines and orthogonal polynomials have been shown to be effective in the modeling of random effects. Applications of RRML to large data sets have shown that reduced rank models are computationally feasible. The robustness and simplicity of splines make them an appealing alternative for modeling of fixed effects. Additionally, the ability of two-dimensional splines to jointly model effects could make them a superior model choice in situations where fixed effect covariates are not completely independent of each other.

LITERATURE CITED

Aarons, L., Baxter, C., and S. Gupta. 2004. Pharmacodynamics of Controlled Release Verapamil in Patients with Hypertension: An Analysis using Spline Functions. Biopharm. Drug Dispos. 25:219-225.

Albuquerque, L. G., and K. Meyer. 2001. Estimates of covariance functions for growth from birth to 630 days of age in Nelore cattle. J. Anim. Sci. 79:2776-2789.

Anderson, S. and B. Peterson. 1996. Growth and food intake curves for group-housed gilts and castrated male pigs. Anim. Sci. 63:457-464.

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.

Bookstein, F. L. 1989. Principle Warps: Thin-Plate Splines and the Decomposition of Deformations. IEEE Trans. Pattern Anal. Mach. Intell. 11:567-585.

Druet, T., Jaffrezic, F., Boichard, D., 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.

Guo, W. 2002. Functional Mixed Effects Models. Biometrics. 58:121-128.

Henderson Jr., C. R. 1982. Analysis of Covariance in the Mixed Model: Higher-Level, Nonhomogeneous, and Random Regressions. Biometrics 38:623-640.

Huisman, A. E., Veerkamp, R. F., and J. A. M. van Arendonk. 2002. Genetic parameters for various random regression models to describe the weight data of pigs. J. Anim. Sci. 80:575-582.

Jamrozik, J., and L. R. Schaeffer. 1997. Estimates of Genetic Parameters with Random Regressions for Yield Traits of First Lactation Holsteins. J. Dairy Sci. 80:762-770.

Jamrozik, J., Schaeffer, L. R, and J. C. M. Dekkers. 1997. Genetic Evaluation of Dairy Cattle Using Test Day Yields and Random Regression Model. J. Dairy Sci. 80:1217-1226.

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 of traits during growth and aging, illustrated with lactation in dairy cattle. Genet. Res. Camb. 64:57-69.

Legarra, A., I. Misztal, and J. K. Bertrand. 2004. Constructing covariance functions for random regression models for growth in Gelbvieh beef cattle. J. Anim. Sci. 82:1564-1571.

Lidauer, M., Mantysaari, E. A., and I. Stranden. 2003. Comparison of test-day models for genetic evaluation of production traits in dairy cattle. Livest. Prod. Sci. 79:73-86

Meinguet, J. 1979. Multivariate Interpolation at Arbitrary Points Made Simple. J. Appl. Math. Phys. 30:292-304.

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. 2001. Estimates of direst an maternal covariance functions for growth of Australian beef calves from birth to weaning. Genet. Sel. Evol. 33:487-514.

Meyer, K. 2004. Scope for a random regression model in genetic evaluation of beef cattle for growth. Livest. Prod. Sci. 86: 69-83.

Molinari, N., Morena, M., Cristol, J. P., and J. P. Daures. 2002. Free knot splines for biochemical data. Comp. Meth. and Programs in Biomedicine. 67:163-167.

Nobre, P. R. C., I. Misztal, S. Tsuruta and J. K. Bertrand. 2003a. Analysis of growth curves of Nellore cattle by multiple-trait and random regression models. J. Anim. Sci. 81:918-926.

Nobre, P. R. C., I. Misztal, S. Tsuruta and J. K. Bertrand. 2003b. Genetic evaluation of growth in Nellore cattle by multiple-trait and random regression models. J. Anim. Sci. 81: 927-932.

Rosenberg, P. S., Katki, H., Swanson, C. A., Brown, L. M., Wacholder, S., and R. Hoover. Quantifying epidemiologic risk factors using non-parametric regression: model selection remains the greatest challenge. Statist. Med. 22:3369-3381.

Scaeffer, L. R., and J. C. M. Dekkers. 1994. Random regressions in animal models for test-day production in dairy cattle. Proceedings of the fifth world congress on genetics applied to livestock production, Guelph. 18:443-446.

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

Strabel, T., and I. Misztal. 1999. Genetic Parameters for First and Second Lactation Milk Yields of Polish Black and White Cattle with Random Regression Test-Day Models. J. Dairy Sci. 82:2805-2810.

Tijani, A., Wiggens, G. R., Van Tassell, C. P., Philpot, J. C., and N. Gengler. 1999. Use of (Co)Variance Functions to Describe (Co)Variances for Test Day Yield. J. Dairy Sci.82:225

Van Der Werf, J. H. J., Goddard, M. E., and K. Meyer. 1998. The Use of Covariance Functions and Random Regressions for Genetic Evaluation of Milk Production Based on Test Day Records. J. Dairy Sci. 81:3300-3308.

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

Wu, H., and J. Zhang. 2002. The study of long-term HIV dynamics using semi-parametric non-linear mixed-effects models. Statist Med. 21:3655-3675.

CHAPTER 3

A PRACTICLE LONGITUDINAL MODEL FOR EVALUATING GROWTH IN GELVIEH CATTLE[1]

---

[1] Robbins, K. R., Misztal, I., and J. K. Bertrand. 2005. *Journal of Animal Science*. 83:29-33. Reprinted here with permission of publisher.

Abstract

Genetic evaluation of growth in Gelbvieh beef cattle was examined by multiple-trait (MTM) and random regression (RRM) analysis. The data set contained 541,108 animals with 1,120,086 records. Approximately 15% of the animals in the dataset had at least one record measured outside of the accepted MTM age ranges for weaning weight (Wwt) and yearling weight (Ywt), and thus were excluded from MTM evaluations. Fourteen percent of Wwt records and 19% of Ywt records were measured outside of the accepted ranges for MTM analysis. Two RRM evaluations were performed using cubic Legendre polynomials (RRML) and linear splines (RRMS) with 3 knots at 1, 205, and 365 days of age. Data set one (d1) utilized all available records, while data set two (d2) included only records measured within MTM ranges (1 d, 160 to 250 d, and 320 to 410 d). The RRML models did not reach convergence until diagonalization was imposed. After diagonalization, it was found that all longitudinal models required fewer iterations to converge than the MTM. Correlations between the MTM, RRML-d2, and RRMS-d2 evaluations were ≥ 0.99 for all three traits, indicating that these models were equivalent when predicting breeding values from data within the MTM age ranges. Correlations between MTM, RRML-d1, and RRMS-d1 were > 0.99 for Bwt and > 0.95 for Wwt and Ywt. The lower correlations for Wwt and Ywt indicate that the added information does affect breeding value prediction. The RRM has the capability to incorporate records measured at all ages into genetic evaluations at a computing cost similar to the MTM.

Key Words: Beef Cattle, Legendre Polynomial, Multiple Trait, Random Regression, Spline

Introduction

Multiple-trait models (MTM) treat all records for a trait as being measured at the same age and having the same parameters. This creates the need for age adjustments, establishment of age ranges, and the subsequent elimination of records measured outside these ranges. These characteristics of MTM have led to the development of random regression models (RRM) for growth traits (Meyer and Hill, 1997; Albuquerque and Meyer, 2001). These models eliminate the need for MTM adjustments by modeling the changes in (co)variance over time. Meyer et al. (2004) showed that these properties of RRM could lead to increases in the accuracy of breeding value prediction.

Although the RRM is theoretically more appealing, previous RRM evaluations using Legendre polynomials (RRML) have yielded inaccurate results (Nobre et al., 2003b). Nobre et al. (2003a) found that estimates of RRML parameters where little data were available were unreliable. Recently Legarra et al. (2004) attempted to eliminate unreliable RRML parameter estimates by deriving them directly from MTM parameters using methods similar to those used by Kirkpatrick et al. (1990; 1994). Bohmanova et al. (2004) looked at RRM for growth traits using simulated data, and found that, in addition to reliable parameters, RRML required diagonalization for numerical accuracy.

While much of the interest in longitudinal models has been focused on RRML, linear splines present a simple alternative. With splines poorly estimated parameters are not a factor as they use MTM parameters. A spline fits a series of functions through control points referred to as knots. The use of linear spline functions can eliminate the numerical problems associated with RRML and decrease the cost of implementation.

The objective of this study was to compare genetic evaluations of growth data in a large beef cattle population using RRM and MTM to determine the practicality of implementing longitudinal evaluations for national data sets.

Materials and Methods

Evaluations were performed on Gelbvieh records spanning 1972-2001. The initial data set contained records on 667,174 animals for birth weight (Bwt), weaning weight (Wwt), and yearling weight (Ywt). All animals that were less than 50 % Gelbvieh, had dams younger than 550 days, or were the only animal in a contemporary group were removed. All records greater than four standard deviations from the mean weight were removed. Seven age-of-dam classes (AOD) were created beginning at 550 days of age. Contemporary groups (CG) included breeder defined CG, sex, and percent Gelbvieh. In addition, Bwt CG included birth year and season. Two data sets were formed. Data set 1 (d1) contained all available records, while data set 2 (d2) contained only the records measured at (1 d, 160 to 250 d, and 320 to 410 d).

The complete data set (d1) contained 1,120,086 records on 541,108 animals. Approximately 15% of the animals had at least one record measured outside the accepted MTM age ranges for Wwt and Ywt. Fourteen percent of Wwt records and 19% of all Ywt records were measured outside the accepted ranges for MTM analysis. However, nearly 65% of records measured outside MTM ranges were within 20 days of age range bounds. A summary of the complete data set is found in Table 3.1.

All parameters used in this study were from work done by Legarra et al. (2004). Legarra et al. (2004) provides detail descriptions of the methods used to estimate all parameters. All models and parameters used in this study were the same as those used by Bohmanova et al.

(2004) with the exception of fixed effect estimation. Since contemporary groups did not remain constant across traits, a continuous regression could not be used in the longitudinal evaluations.

Fixed effects for all models were determined using the following equation:

$$fixed_{ijlt} = cg_i + \alpha_{lt}(age_l - age_t) + \alpha_{lt}(age_l - age_t)^2 + \alpha_{lt}(age_l - age_t)^3 + aod_j,$$

where cg_i = contemporary group i; α_{lt} = linear, quadratic, and cubic regression coefficients at age l and trait t; age_l = age l of animal; age_t = equals the reference age of trait t; aod_j = age of dam class j. In longitudinal models the equation was nested in three dummy variables, representing the three traits. In addition, the age of dam classes and contemporary groups were re-numbered for Wwt and Ywt traits. These modifications were made to ensure that the equation was equivalent for all models used in this study.

The MTM model presented in scalar notation was:

$$y_{ijklmnt} = fixed_{ijlt} + dir_{kt} + mat_{mt} + mpe_{mt} + e_{ijklmnt},$$

where $y_{ijklmnt}$ = weight for trait t at age l of contemporary group i, and age of dam group j; dir_{kt} = the random direct additive effect of animal k for trait t; mat_{mt} = the random maternal effect of dam m for trait t; mpe_{mt} the maternal permanent environmental effect of dam m for trait t; and $e_{ijklmnt}$ = random residual effect. Direct and maternal effects were assumed correlated.

The RRM using linear splines in scalar notation was:

$$y_{ijklmn} = fixed_{ijl} + \sum_{d=0}^{2} s_{dl}dir_{dk} + \sum_{d=0}^{2} s_{dl}pe_{dk} + \sum_{d=0}^{2} s_{dl}mat_{dm} + \sum_{d=0}^{2} s_{dl}mpe_{dm} + e_{ijklmn},$$

where y_{ijklmn} = weight for trait t of contemporary group i, and age of dam group j; dir_{dk} and pe_{dk} = spline coefficients d for additive direct and permanent environmental effects for animal k; mat_{dm} and mpe_{dm} = spline coefficients d for maternal and maternal permanent environmental effects for dam m; e_{ijklmn}= weighted heterogeneous random residual; and s_{dl} = dth coefficient of the linear spline function for an observation taken at age l.

The RRM, constructed with cubic Legendre polynomials, was defined as:

$$y_{ijklmn} = fixed_{ijl} + \sum_{d=0}^{3} z_{dl} dir_{dk} + \sum_{d=0}^{3} z_{dl} pe_{dk} + \sum_{d=0}^{3} z_{dl} mat_{dm} + \sum_{d=0}^{3} z_{dl} mpe_{dm} + e_{ijklmn},$$

where y_{ijklmn} = weight for trait t of contemporary group i, and age of dam group j; dir_{dk} and pe_{dk} = random regression coefficients d for additive direct and permanent environmental effects for animal k; m_{dm} and mpe_{dm} = random regression coefficients for maternal and maternal permanent environmental effects for dam m; e_{ijklmn}= weighted heterogeneous random residual ; and z_{dl} = dth coefficient of Legendre polynomial for observation taken at age l.

The weights for the heterogeneous residual variance were modeled using linear splines, as in Bohmanova et al. (2004) and Legarra et al. (2004), and implemented by weighting each observation. Solutions were computed by the program BLUP90IOD, which uses iteration on data with the precondition conjugate gradient iteration (Tsuruta et al., 2001).

Results and Discussion

In this study, MTM was fit to d2 only, while RRML and RRM with linear splines (RRMS) were fit to d1 and d2. The number of iterations required for convergence of each model is reported in Table 3.2. Both the RRML and RRMS required fewer iterations than the MTM, indicating that longitudinal models do not have higher computational requirements than multiple trait models. This is in agreement with earlier results showing that similar RRMs required less memory and time than MTM (Nobre et al., 2003b). However, it should be noted that only the RRML was diagonalized, as convergence could not be reached otherwise. Bohmanova et al. (2004) obtained convergence of RRML without diagonalization using a smaller simulated data set. In that case, the diagonalization decreased computing over five times. In order to achieve the most accurate results, a strict convergence criterion of 10^{-14} was assigned to all models. Previous work by Nobre et al. (2003b) found that strict convergence criteria were necessary to obtain accurate RRML results. Similar strict convergence criterion for the preconditioned conjugate gradient iteration was necessary in a study by Tsuruta et al. (2001).

Table 3.3 shows Pearson correlations of fixed effects between the MTM and longitudinal models when applied to data set two. The estimation of fixed effects was equivalent for all models used in this study. Correlations for direct effects between all models can be found in Table 3.4. As expected, correlations for Bwt were the highest as there are no age ranges for this measurement. Due to the longitudinal models' ability to incorporate changes in (co)variance, the correlations were lower for Wwt and Ywt. However, these decreases were relatively small in correlations that involved models utilizing data set two. This suggests that the changes in variance within the MTM ranges are small. The decrease in correlations of RRML-d1 and RRMS-d1 with MTM for Wwt and Ywt were larger, indicating that the inclusion of all available records does affect BV predictions. However, due to the low numbers of outlying measurements in the data set these correlations were still > 0.95. Inclusion of large amounts of data measured outside the MTM age ranges could result in even lower correlations.

Correlations for maternal effects are found in Table 3.5. Some discrepancy exists between the RRML-d2 and RRMS-d2 for weaning and yearling maternal effects. The high correlations for Bwt were expected, as the models are equivalent when estimating effects at birth. The RRML, however, could have some trouble fitting a function for maternal effects at weaning due to large biological differences between Bwt and Wwt maternal effects, resulting in lower correlations for Wwt maternal effects. Birth maternal effects are largely a measure of amniotic effects, while weaning maternal effects are a measure of milking. The lower correlations for both models at yearling age are relatively unimportant as maternal effects post weaning are merely residuals of earlier maternal effects. Albuquerque and Meyer (2001) found that heritabilities for maternal effects peak around 110 to 120 days of age, meaning higher responses would be expected if selection on maternal effects in this age range were practiced. Unlike the RRML, the current MTM does not predict maternal BV at this age. Both the RRMS-d2 and RRML-d2 had high correlations with the MTM for maternal permanent environmental (MPE) effects at all ages. The correlations, reported in Table 3.6, indicate that the longitudinal models and the MTM are very similar in prediction of MPE effects.

In Table 3.7 sire rank correlations between MTM and the longitudinal models are presented. These correlations are for sires with 50 or more progeny records for at least one of the three traits. The high rank correlations indicate that implementing a longitudinal evaluation will have little impact on the rank of moderate to high accuracy sires with greater than 50 progeny when the models use only information measured within the usual MTM age ranges. However, lower rank correlations between RRML-d1 and RRMS-d1 with MTM were observed for Wwt and Ywt. Table 8 shows that when all available data are used the number of progeny records increase considerably for some sires. Furthermore many new sires are generated by the inclusion of all available records. Table 9 shows that the increase in records does have an effect on predicted breeding values, especially Wwt breeding values. These factors contribute to the lower sire rank correlations when all available data is used.

Although no age restrictions were placed on d1, the majority of records in d1 were close to the MTM age ranges. While the RRML and RRMS evaluations were similar when using d1, the models may not agree when large numbers of early Wwt and late Ywt records are present. A 3-knot spline was used in RRMS, the third knot being at 365 days. As a result, the variance of records measured after 365 days of age was determined using an extension of the linear function between weaning and yearling knots. As an animal's age increases beyond 365 days, the interpolation becomes less accurate. Bohmanova et al. (2004) compared variances between RRML and RRMS. While variances and correlations were quite similar for points between Wwt and Ywt, the variances as approximated by RRMS around 100 d were visibly smaller (and correlations larger) than those obtained from the RRML, and the reverse was true for observations over 365 d. Adding a knot at 100 d strongly decreased the differences in variances and correlations. When many records are available around 100 d and over 400 d, it may be useful to add extra knots to RRMS. Furthermore, as age increases, growth approaches an asymptotic value that, in addition to little data being available, can cause high order polynomial regressions like those used by RRML to become erratic.

Implications

The results of this study indicate that longitudinal models can be implemented effectively in beef cattle growth evaluations. The simplicity and easy implementation of the linear spline random regression model makes it an appealing alternative to the current multiple-trait model. The random regression model utilizing cubic Legendre polynomials requires diagonalization, however, it allows for smoother (co)variance functions. Both random regression models give practical and more flexible evaluations, while providing a more theoretically sound alternative to the multiple-trait model with relatively small cost of implementation.

Literature Cited

Albuquerque, L. G., and K. Meyer. 2001. Estimates of covariance functions for growth from birth to 630 days of age in Nelore cattle. J. Anim. Sci. 79:2776-2789

Bohmanova, J., I. Misztal, and J. K. Bertrand. 2004. Studies on multiple trait and random regression models for genetic evaluation of beef cattle for growth. J. Anim. Sci. (In press).

Legarra, A., I. Misztal, and J. K. Bertrand. 2004. Constructing covariance functions for random regression models for growth in Gelbvieh beef cattle. J. Anim. Sci. 82:1564-1571.

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 of traits during growth and aging, illustrated with lactation in dairy cattle. Genet. Res. Camb. 64:57-69

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. 2004. Scope for a random regression model in genetic evaluation of beef cattle for growth. Livest. Prod. Sci. 86: 69-83.

Nobre, P. R. C., I. Misztal, S. Tsuruta and J. K. Bertrand. 2003a. Analysis of growth curves of Nellore cattle by multiple-trait and random regression models. J. Anim. Sci. 81:918-926.

Nobre, P. R. C., I. Misztal, S. Tsuruta and J. K. Bertrand. 2003b. Genetic evaluation of growth in Nellore cattle by multiple-trait and random regression models. J. Anim. Sci. 81: 927-932.

Tsuruta, S., I. Misztal, and I. Stranden. 2001. Use of the preconditioned conjugate gradient algorithm as a generic solver for mixed model equations in animal breeding applications. J. Anim. Sci. 79:1166-1172.

Table 3.1. Characteristics of the data set

			Weight, kg				Age, d		
Traits[a]	Records	Sires	Mean	SD	Min	Max	Mean	Min	Max
Bwt	508,918	16,786	38.9	5.3	19.1	59.0	-	-	-
Wwt	460,346	16,069	258.3	45.8	75.3	440.9	202.4	76	328
Ywt	150,822	9,743	449.7	81.2	162.8	772.0	357.84	216	501

[a] Bwt = birth weight. Wwt = weaning weight. Ywt = yearling weight.

Table 3.2. Convergence in iterations

Model[a]	Iterations
MTM	678
RRML-d1	456
RRML-d2	274
RRMS-d1	556
RRMS-d2	254

[a] MTM = multiple trait model. RRML = random regression model utilizing Legendre polynomials. RRMS = random regression utilizing linear splines. d1= complete data set. d2= data set containing only records measured within MTM age ranges.

Table 3.3. Correlations of fixed effects between multiple trait and longitudinal models

Effect[a]	Birth Weight	Weaning Weight	Yearling Weight
CG	0.999	0.998	0.999
AOD	0.999	0.999	0.999

[a] CG = contemporary group. AOD = age of dam class.

Table 3.4. Correlations for direct additive effects

Model[a]	Birth Weight		Weaning Weight		Yearling Weight	
			Data Set 2[b]			
	RRML	MTM				
RRMS	0.999	0.999	0.998	0.992	0.996	0.992
RRML		0.998		0.992		0.991
			Data Set 1			
	RRML	MTM				
RRMS	0.999	0.994	0.996	0.953	0.993	0.951
RRML		0.990		0.959		0.956

[a] MTM = multiple trait model. RRML = random regression model utilizing Legendre polynomials. RRMS = random regression utilizing linear splines. [b] Data set 1= complete data set. Data set 2= data set containing only records measured within MTM age ranges.

Table 3.5. Correlations of maternal effects between multiple trait and longitudinal models

Model[a]	Birth Weight	Weaning Weight	Yearling Weight
RRMS-d2	0.995	0.988	0.970
RRML-d2	0.991	0.975	0.981

[a] RRML = random regression model utilizing Legendre polynomials. RRMS = random regression utilizing linear splines. d2= data set containing only records measured within MTM age ranges.

Table 3.6. Correlations of maternal permanent environmental effects between multiple-trait and longitudinal models

Model[a]	Birth Weight	Weaning Weight	Yearling Weight
RRMS-d2	0.984	0.985	0.986
RRML-d2	0.983	0.985	0.986

[a] RRML = random regression model utilizing Legendre polynomials. RRMS = random regression utilizing linear splines. d2= data set containing only records measured within MTM age ranges.

Table 3.7. Rank correlations of sires with greater than 50 progeny between Multiple trait and longitudinal models

Model[a]	Birth Weight	Weaning Weight	Yearling Weight
RRMS-d2	0.999	0.992	0.990
RRML-d2	0.999	0.993	0.987
RRMS-d1	0.996	0.961	0.956
RRML-d1	0.995	0.965	0.948

---

[a] RRML = random regression model utilizing Legendre polynomials. RRMS = random regression utilizing linear splines. d1= complete data set. d2= data set containing only records measured within MTM age ranges.

Table 3.8. Increases in sire information when all available data are utilized versus MT data

	No. Sires[a]	
Increase[b]	Wwt	Ywt
New sires	626	648
> 0%	9,008	4,898
≥ 100%	655	545
≥ 500%	31	38

[a]Increase = the increase in progeny records due to the utilization of all available data. [a]Bwt = birth weight. Wwt = weaning weight. Ywt = yearling weight. [b]New sires = animals that had no progeny records in the edited data set, but had at least one progeny record in the complete data set

Table 3.9. Correlations of sires that gain greater than 100% more progeny records when all data was utilized

Model[a]	Weaning Weight	Yearling Weight
	MTM	MTM
RRMS-d1	0.876	0.927
RRML-d1	0.884	0.919

[a]MTM = multiple trait model. RRML = random regression model utilizing cubic Legendre polynomials. RRMS = Random regression model utilizing linear splines. d1= complete data set. d2= data set containing only records measured within MTM age ranges.

CHAPTER 4

JOINT LONGITUDINAL MODELING OF AGE OF DAM AND AGE OF ANIMAL[1]

---

[1] Robbins, K. R., Misztal, I., J. K. Bertrand. To be submitted to the *Journal of Animal Science*.

Abstract

We examined the joint modeling of age of dam (AOD) and age of animal in a random regression model analysis of growth in Gelbvieh cattle. The first method (M1) was analogous to the multiple trait analysis and consisted of AOD as a nested class variable and a cubic polynomial regression on age nested within birth, weaning and yearly weights. The second method (M2) utilized two-dimensional linear splines, with age knots at 150 d, 205 d, 270 d, 340 d, and 390 d. The AOD knots were placed at 725 d, 1464 d, and 2189 d. A data set containing Gelbvieh growth records was split along contemporary groups into two data sets. Data set 1 (D1) contained 316,078 records and was used for prediction by mixed model equations. Data set 2 (D2) contained 164,167 records and was used for cross validation. Models were evaluated based on R^2, average squared error (ASE), percent bias, and plots of solutions. ASE for weights associated with birth weight, weaning weight and yearling weight for M1 were 15, 505, and 703 kg^2. Due to the nonlinearity of M2, large jumps in fixed effect estimates were observed outside the two-dimensional grid. To eliminate this problem, weighted one-dimensional splines were used for extrapolation beyond the two-dimensional grid. For M2, with hybrid spline extrapolation, the ASE were 15, 542, and 777 kg^2. Creation of optimal two-dimensional splines is difficult when data is clustered.

Key Words: polynomial regression, two-dimensional spline, cross validation.

Introduction

While much work has been done on the modeling of random effects in random regression models (RRM) for growth, relatively little attention has been paid to the modeling of fixed effects in this context. When evaluating growth, two continuous covariates are commonly modeled, the age of dam effect (AOD) and the animal's age. Multiple trait models (MTM) are currently used to analyze growth in beef cattle and as a result, records tend to be clustered around birth, weaning and yearling ages. With such age distributions, polynomial functions nested within each trait seem to be an obvious choice for the modeling of these effects. However, as RRM models become more widely accepted, larger and more continuous age ranges will be recorded for evaluation. This could make nesting more difficult and outlying age records could make polynomials vulnerable to artifacts.

An alternative to polynomial regressions are splines. Splines are a series of polynomial functions fit through control points, referred to as knots. It has been shown that spline functions are resistant to artifacts (Aarons et al., 2004; Druet et al., 2003). Unlike polynomial regressions in which a small subset of data can affect the entire function, splines are defined by a series of polynomials that are affected only by their bounding knots (Molinari et al., 2002). While one-dimensional splines provide a more robust model, they still require nesting when modeling AOD and age of animal. Two-dimensional splines can provide a generalized and robust model for fixed effects. The selected two dimensional knots provide automatic nesting and implicit modeling of interactions while reducing the effects of outlying records.

The purpose of this cross validation study was to compare the performance of several models in evaluating growth in beef cattle, and provide a basic methodology for fitting two-dimensional splines to biological data.

Materials and methods

Two methods were employed to model AOD and age of animal. Method 1 (M1) modeled AOD as a within-trait nested class variable and utilized a within-trait nested cubic polynomial regression to model age of animal. Method 2 (M2) jointly modeled AOD and age of animal with a two-dimensional spline. All evaluations were performed on the data set used by Robbins et al. (2005). The data set was split within contemporary groups (CG). All CG with less than 15 animals were eliminated to ensure accurate estimation of CG by the prediction data set.. Data set 1 (D1) contained 316,078 growth records on Gelbvieh cattle and was used for mixed model evaluation. Data set 2 (D2) contained 164,167 records and was used for cross validation. Solutions obtained by the program BLUP90IOD (Tsuruta et al., 2001) from the analysis of D1 were used to predict the records of animals in D2. Using the actual and predicted D2 records, the average squared errors (ASE) and percent bias were computed for each model at each trait (birth weight, Bwt; weaning weight, Wwt; and yearling weight, Ywt). The ASE, percent bias, R^2, and plots of fixed effect solutions were used to evaluate each model.

All random effects were modeled using linear splines. The equation for random effects in scalar notation was:

$$random_{ijklm} = \sum_{d=0}^{2} s_{dl} dir_{dk} + \sum_{d=0}^{2} s_{dl} pe_{dk} + \sum_{d=0}^{2} s_{dl} mat_{dm} + \sum_{d=0}^{2} s_{dl} mpe_{dm} + e_{ijklm},$$

where $random_{ijklm}$ = sum of random effects for trait t, and age of dam group j; dir_{dk} and pe_{dk} = spline coefficients d for additive direct and permanent environmental effects for animal k; mat_{dm} and mpe_{dm} = spline coefficients d for maternal and maternal permanent environmental effects for dam m; e_{ijklm}= weighted heterogeneous random residual; and s_{dl} = dth coefficient of the linear spline function for an observation taken at age l. This is the same equation as used by Robbins et al. (2005).

The fixed effect model using within-trait nested cubic polynomial regressions on age and within-trait nested AOD classes was:

$$fixed_{ijlt} = cg_i + \alpha_{lt}(age_l - age_t) + \alpha_{lt}(age_l - age_t)^2 + \alpha_{lt}(age_l - age_t)^3 + AOD_{jt},$$

where cg_i = contemporary group i, comprised of animals of the same sex, percent Gelbvieh, and from the same breeder defined management groups; α_{lt} = linear, quadratic, and cubic regression coefficients at age l and nested in trait t; age_l = age l of animal; age_t = the reference age of trait t; AOD_j = age of dam class j nested in trait t. The age of dam classes were renumbered for Wwt and Ywt traits for nesting purposes.

A second model that contained the same within-trait fixed effects plus an additional AOD by age of animal interaction was fit to the data and is described below:

$$fixed_{ijlt} = cg_i + \alpha_{lt}(age_l - age_t) + \alpha_{lt}(age_l - age_t)^2 + \alpha_{lt}(age_l - age_t)^3 + aod_j + \alpha_{lt}(AOD_j * age_l)$$

where $AOD_j{*}age_l$ = the interaction of AOD class j by age of animal 1.

The two dimensional spline model can be written as:

$$fixed_{ijl} = cg_i + cf_{lj} * \alpha_{lj} + cf_{l(j-1)} * \alpha_{1(j-1)} + cf_{(l-1)j} * \alpha_{(l-1)j} + cf_{(l-1)(j-1)} * \alpha_{(l-1)(j-1)}$$

where c*f*= the coefficient of an animal with age and AOD such that: $age_{(l-1)} < age \leq age_l$ and $AOD_{(j-1)} < AOD \leq AOD_j$; and α_{lj} = the estimated knot value for age l and AODj.

The coefficients for the two dimensional splines were determined as:

$$cf_{lj} = \sqrt{x^2 + y^2}$$

where x is 1- the distance of the age of animal from the knot for age of animal; y is determined by 1- the distance of the AOD from the knot for AOD when 0<=distance<=1.

Since the two-dimensional spline was poor at extrapolation beyond the two-dimensional grid, a model was run that used the weighted sum of 1-dimesional splines for extrapolation beyond the grid knots. The equation was:

$$\sum_{i=1}^{nk} w_i * lcf \quad * knot_{ad}$$

where w_i = weighting factor for 1-dimensional $spline_i$; lcf = the linear spline coefficient a for 1-dimesional spline extrapolation; $knot_{ad}$ = the two dimensional spline knot for age of animal a and age of dam d; nk = the number of one-dimensional spline functions. An alternative is to create a function with an asymptote such that $knot_{ad}$ = a one-dimensional spline knot at age a, with age a being 1day beyond the bound of the two-dimensional grid, and age of dam d, with age of dam d being 1 day beyond the bound of the two dimensional grid. This results in a non-continuous model with an increased number of knots.

Since evaluation models were over parameterized, mean squared error could not be used as there were no degrees of freedom. Therefore, the ASE was used to evaluate the fixed effect models. The ASE was computed as:

$$ASE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \Big/ n$$

where y_i = the weight of animal i; $\hat{y}_i$ = the predicted record of animal i; n = the number of records contained in the test data set. In addition to ASE, percent bias was calculated as:

$$percent\ bias = [\sum_{i=1}^{n} (y_i - \hat{y}_i) \Big/ \sum_{i=1}^{n} y_i \ ]*100$$

The steps involved in model selection differ greatly between M1 and M2. These differences represent an important distinction between model types and should be taken in to consideration when formulating a fixed effect model. To construct the polynomial model one needs to consider the order of the polynomial as well as the need and location for nesting. Since AOD was modeled as categorical, class boundaries had to be defined. Such considerations make model selection a relatively simple process for M1.

Creating an optimal two-dimensional spline model can be much more time consuming than with polynomial regressions. Splines are approximations that depend heavily on the location of the knots. There must be enough knots to adequately model the shape of the function, and there must be enough records in each interval between knots to accurately estimate knot values. Unfortunately, there is no automatic procedure for the selection of knots; however, there are some general rules that can aid this process. Wold (1974) suggests the use of as few knots as possible, no more than one extremum and one inflection point per interval, and the location of knots close to inflection points. In the case of the two dimensional spline, the application of these rules to each variable separately can provide a good starting point. In addition, the two-dimensional spline's inability to model data outside the two-dimensional grid necessitates the placement of knots at extreme values. However, if data is sparse around the extremums, the use of weighted spline extrapolation may give the best results.

Once a base model has been established there are some generalized procedures for the addition of knots to the model. One procedure is to place an additional knot at the median of the existing interval (Rosenburg et al., 2003). This process could be useful when the data is continuously distributed. In the case of growth data in beef cattle, both age and AOD are clustered, thus limiting the areas in which knots can be placed. In such a case the median may not be the best place to add additional knots, however, the general principles of this procedure can be useful in expanding the base model. When dealing with disjointed data placing knots at the end points of each cluster may be a good idea, however, if data is sparse at the endpoints placing the knots closer to the center can provide better results.

When using two-dimensional splines, variables may behave differently depending on the value of another variable. In such a case placing knots based on each variable's curve alone may not be optimal. To account for possible interactions or nesting effects, conditional plots can be of value. Plotting a variable by each interval of the other variable can help in determining how the two variables interact. In such a case, it is best to place as few knots as possible that allow enough flexibility to model possible interactions and nesting effects. It is important to remember, that while an optimized spline is robust against artifacts, the flexibility of the spline model makes it highly susceptible to artifacts when knots are poorly placed (Wold, 1974).

Results and Discussion

Cross validation results in Table 4.1 show that M1 performed well. The model containing AOD by age of animal interactions did have lower R^2 values than models without the interaction effect, suggesting no interaction is present in the data. The interaction model showed increases in ASE and negative biases for Wwt and Ywt. The relatively large and negative percent bias values show that the interaction model is over predicting records, this is likely due to over fitting of the model to D1. It was expected that the nested polynomials would perform well given the disjointed nature of the age distributions. The fact that these data were collected for multiple trait evaluation provides obvious nesting choices. The distinct Bwt, Wwt, and Ywt groups, coupled with the high density of records within each group, makes nested polynomials an appealing model choice. However, as RRM become more widely used, distinctions between nesting groups will become blurred, and areas of sparse, outlying data could affect the polynomial regressions.

The fixed effect plot for M1 can be seen in Figure 4.1. The function on age shows a period of steep linear growth between 75 d and 275 d, followed by a period of declining slope between 275 and 325 days of age. The declining slope is the result of the non-continuity of the nested regressions. The point at 275 d is calculated with the regression nested in Wwt while the point at 325 d is calculated by the regression nested in Ywt. Clearly the two nested functions are disjointed. Beyond 325 d another period of steep linear growth can be seen. When looking at the AOD functions at birth and yearling in figure 4.2, there is a period of linear increase until 1500 d followed by a shallow incline to a relatively flat plateau at 2200 d. At yearling age the AOD function shows a linear incline to at plateau at 1500 d.

For this application, the best fitting two dimensional spline models contained 5 age knots at 150, 205, 270, 340, and 390 days of age, with birth being analyzed separately. The function below 150 d was modeled as a decreasing function from 150 d towards zero. The function above 390 d was modeled as an increasing function. Models that contained more age knots were erratic and seemed to be influenced by artifacts, while models with fewer knots performed poorly in cross validation. For AOD, three knots were placed at 725, 1464, and 2190 d. The AOD function was much more sensitive to artifacts than age of animal and thus only a few knots could be used. As AOD increases, data becomes considerably sparser. It was found that forcing a flat function through later dam ages at birth, weaning, and yearling yielded the best results. It should be noted that there is some flexibility in choosing knots, as several combinations of knots yielded comparable results to the above noted model. The exclusion of Bwt in the two-dimensional spline models had only a small effect on ASE and percent bias, but did provide the best model.

As seen in Table 4.1, M2 performed well with the extended grid and weighted spline extrapolation methods. The parity of M1 and M2 at birth would be expected as there is no age variation, with the modeling of AOD effects being the only difference between models. Though there are some differences in R^2, ASE, and percent bias at Wwt and Ywt, M1 and M2 had similar fits for these traits. These results suggest that M2 is capable of automatically nesting AOD within age of animal, but does not provide a superior fit to the data. While M1 does have seven additional fixed effect parameters than M2, their effect on model complexity is negligible when weighted against the over 16,000 contemporary groupings.

When looking at the graph of M2 in Figure 4.3a, it appears that utilizing two-dimensional functions to extrapolate beyond the grid can result in large jumps in the estimated effects. This results from the fact that, unlike one-dimensional splines, two-dimensional coefficients must be forced to sum to a constant. Once outside the grid this restriction is removed and knot coefficients suddenly jump to values that no longer sum to this constant. As seen in figure 4.3b, the use of weighted spline extrapolation greatly alleviates this problem. The weighted spline function allows the sum of knot values to gradually increase or decrease from one. As well as giving smoother graphs, the use of the weighted interpolation gives lower ASE and percent bias as shown in Table 4.1. Another solution to this problem is the extension of the two-dimensional grid to encompass all data. This method performs well in terms of ASE and percent bias, but graphs of solutions in Figures 4.3c show it can be subject to artifacts. An alternative to the previously described M2 methods involves the elimination of the bounding knot farthest from any given data point. This would leave only three knots, allowing for a linear formulation of knot coefficients. In addition to this simplified triangular methodology, the inclusion of both fixed and random interaction effects could be effective for modeling of data sets containing multiple growth curves.

The graph of M2, found in Figure 4.3b, shows an almost linear growth throughout the function with some curvature present. This linear growth is also observed in each of the two polynomial regressions found in Figure 4.1, however, with M2 the function is continuous. This property of M2 could make it a better choice for modeling data distributed continuously across all ages. The graph of the AOD effect at birth, found in Figure 4.2, shows a linear incline to a flat plateau. When looking at AOD curves for weaning and yearling, slowly increasing functions can be seen prior to 1300 d followed by steeper linear inclines that plateau. For birth and weaning these plateaus are reached at 2200 d as do the M1 AOD functions. For yearling, M2 curves reach a plateau at around 1500 d much like M1. This is not surprising as AOD has little effect after weaning and AOD curves tend to be flatter. The AOD curves seemed particularly sensitive to artifacts when more than 3 or 4 knots were used. Such curves were very erratic and yielded poor results in cross validation. As can be seen in Figure 4.2, estimates of AOD effects obtained from M1 are of a larger magnitude than estimates obtain from M2. This is due to the location of the overall mean in each model. In M2, the overall mean was contained in contemporary group, and as a result, is not present in the AOD and age graphs. With the polynomial regression, some of the overall mean is present in the cross classified AOD effect, and therefore is present in the AOD and age graphs. The presence and absence of this mean has an impact on the magnitude of age and AOD estimates. However, putting scale issues aside, the graphs of M1 and M2 are similar for age of animal and AOD.

The clustering of data around birth, 205 d, and 365 d makes the use of nested polynomials a relatively simple and effective way to model fixed effects in this application. However, as the use of RRM models becomes a more standard practice, the nesting of polynomials will become increasingly difficult as collection of data across ages becomes more continuous. Due to the disjoint nature of the nested polynomials, evaluation of animals with records located between Wwt and Ywt age ranges could be problematic.

Given the current state of the industry in which records are clustered within predefined age ranges, M2 does not have an advantage over traditional polynomial regressions. Previous applications of two-dimensional splines have been in the form of thin plate splines used in the context of engineering and graphical applications (Meinguet, 1979). In such instances, data is collected in a grid like manner or such that observations are located at key points given a known three-dimensional shape Bookstein (1989). Under such conditions, thin plate splines are very effective; however, in the present application neither of these conditions is met. This does not mean that two-dimensional splines cannot be effective, but they may not provide optimal performance. Despite this, the polynomial regression's potential susceptibility to artifacts could make two-dimensional splines a more attractive choice.

Implications

While nested polynomials perform well, their sensitivity to artifacts and need for nesting could create problems as records are measured for increasingly wide age ranges. Although two-dimensional splines do not have superior performance with clustered data, their automatic nesting and robustness could make them an appealing choice for data sets in which nesting is difficult and areas of sparse outlying data are present. When modeling data collected for longitudinal analysis, the continuous nature of the two-dimensional spline may yield superior performance relative to the disjointed, nested polynomial model.

Literature Cited

Aarons, L., Baxter, C., and S. Gupta. 2004. Pharmacodynamics of Controlled Release Verapamil in Patients with Hypertension: An Analysis using Spline Functions. Biopharm. Drug Dispos. 25:219-225.

Bookstein, F. L. 1989. Principle Warps: Thin-Plate Splines and the Decomposition of Deformations. IEEE Trans. Pattern Anal. Mach. Intell. 11:567-585.

Druet, T., Jaffrezic, F., Boichard, D., 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.

Meinguet, J. 1979. Multivariate Interpolation at Arbitrary Points Made Simple. J. Appl. Math. Phys. 30:292-304.

Molinari, N., Morena, M., Cristol, J. P., and J. P. Daures. 2002. Free knot splines for biochemical data. Comp. Meth. and Programs in Biomedicine. 67:163-167.

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

Rosenberg, P. S., Katki, H., Swanson, C. A., Brown, L. M., Wacholder, S., and R. Hoover. Quantifying epidemiologic risk factors using non-parametric regression: model selection remains the greatest challenge. Statist. Med. 22:3369-3381.

Tsuruta, S., I. Misztal, and I. Stranden. 2001. Use of the preconditioned conjugate gradient algorithm as a generic solver for mixed model equations in animal breeding applications. J. Anim. Sci. 79:1166-1172.

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

Appendix 4.1

An example of the calculations for the weights (w_i) and spline coefficients (cf) utilized for extrapolation with weighted splines. The w_i and cf for an animal's record taken at 216d of age with a 3000d old dam is computed. Assuming the model used in this study, the bounding age knots for this record are at 205d and 270d with the last age of dam knot place at 2190d. If the extrapolation function is modeled as decreasing beyond age of dam 2190 the w_i and cf would be calculated as follows:

$w_1 = 1-[(216-205)/(270-205)] = .83$

$w_2 = [(216-205)/(270-205)] = .17$

cf = (2190/3000) = .73

Weighted spline extrapolation = $.83*.73*\text{knot}_{(205,2190)} + .17*.73*\text{knot}_{(270,2190)}$

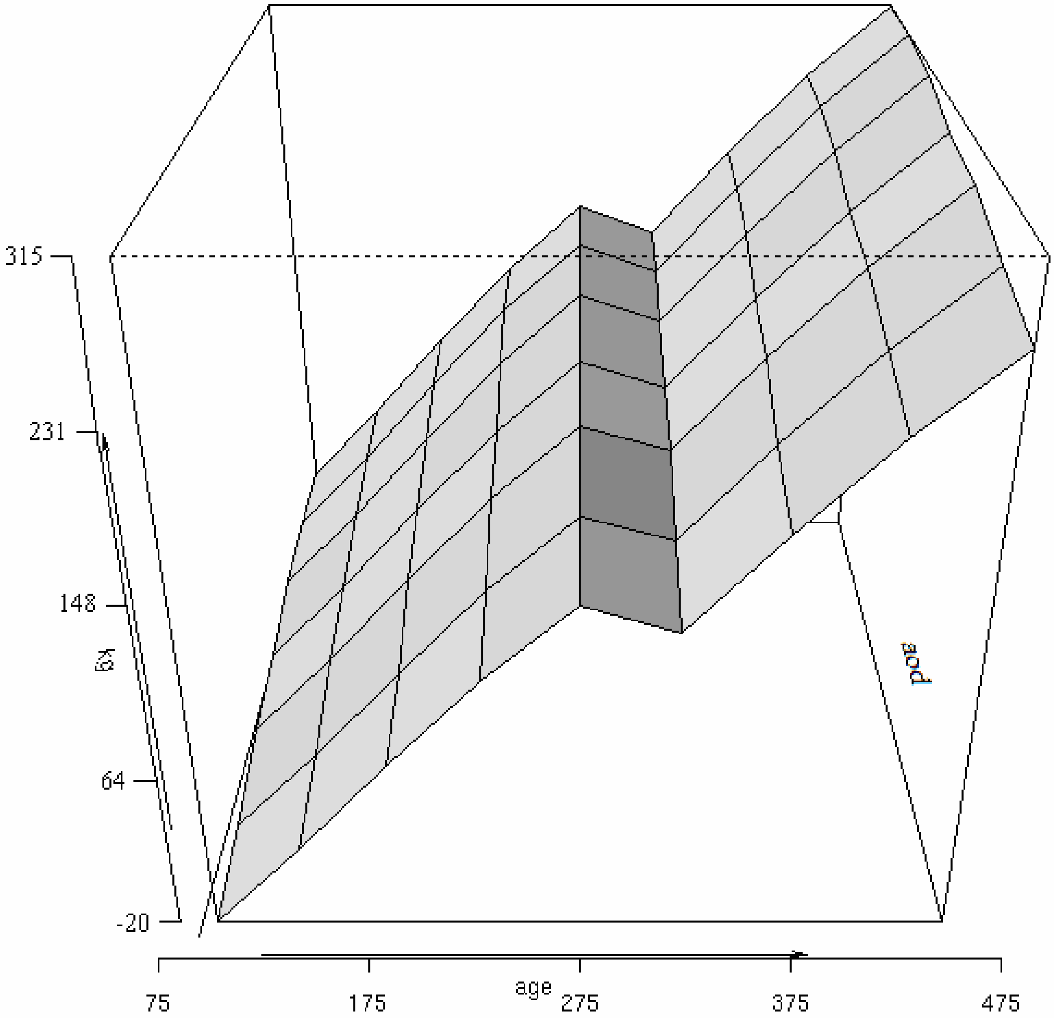
where $\text{knot}_{(205,2190)}$ is the two-dimensional spline knot, as estimated by the mixed model equations, at 205d of age and a dam age of 2190d. $\text{Knot}_{(270,2190))}$ is the two-dimensional spline knot at 270d of age and a dam age of 2190d.

Table 4.1. Cross validation results

Measurement	Birth Weight	Weaning Weight	Yearling Weight
M1			
R^2	.40	.74	.87
ASE	15 kg^2	505 kg^2	703 kg^2
%bias	.56	-0.14	0.10
M1 with interaction			
R^2	.40	.72	.85
ASE	15 kg^2	545 kg^2	837 kg^2
%bias	.56	-2.34	-2.53
M2			
R^2	.40	.70	.84
ASE	15 kg^2	600 kg^2	876 kg^2
%bias	1.12	1.31	1.12
M2 with weighted spline extrapolation			
R^2	.40	.73	.86
ASE	15 kg^2	542 kg^2	777 kg^2
%bias	.74	.36	0.52
M2 with extended grid			
R^2	.40	.72	.86
ASE	15 kg^2	547 kg^2	802 kg^2
%bias	.87	.52	0.65

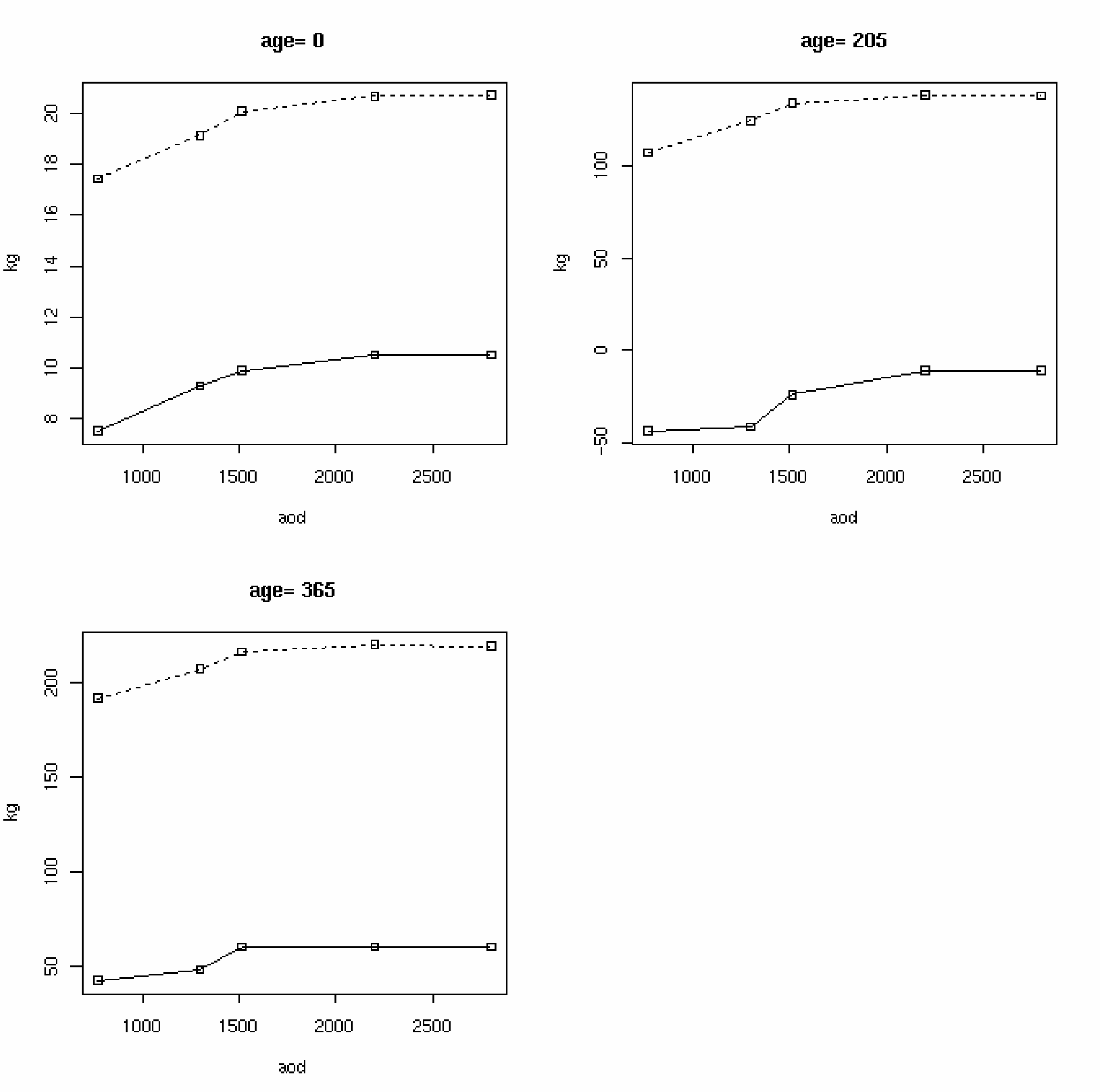
[a] ASE = the average squared error. M1= polynomial regression on age nested in trait an age of dam classes nested in trait. M2= Two-dimensional spline.

Figure 4.1. 3-d plots of Polynomial regression model nested in trait [a]



[a] There are no observations plotted between 275 d and 325 d of age.

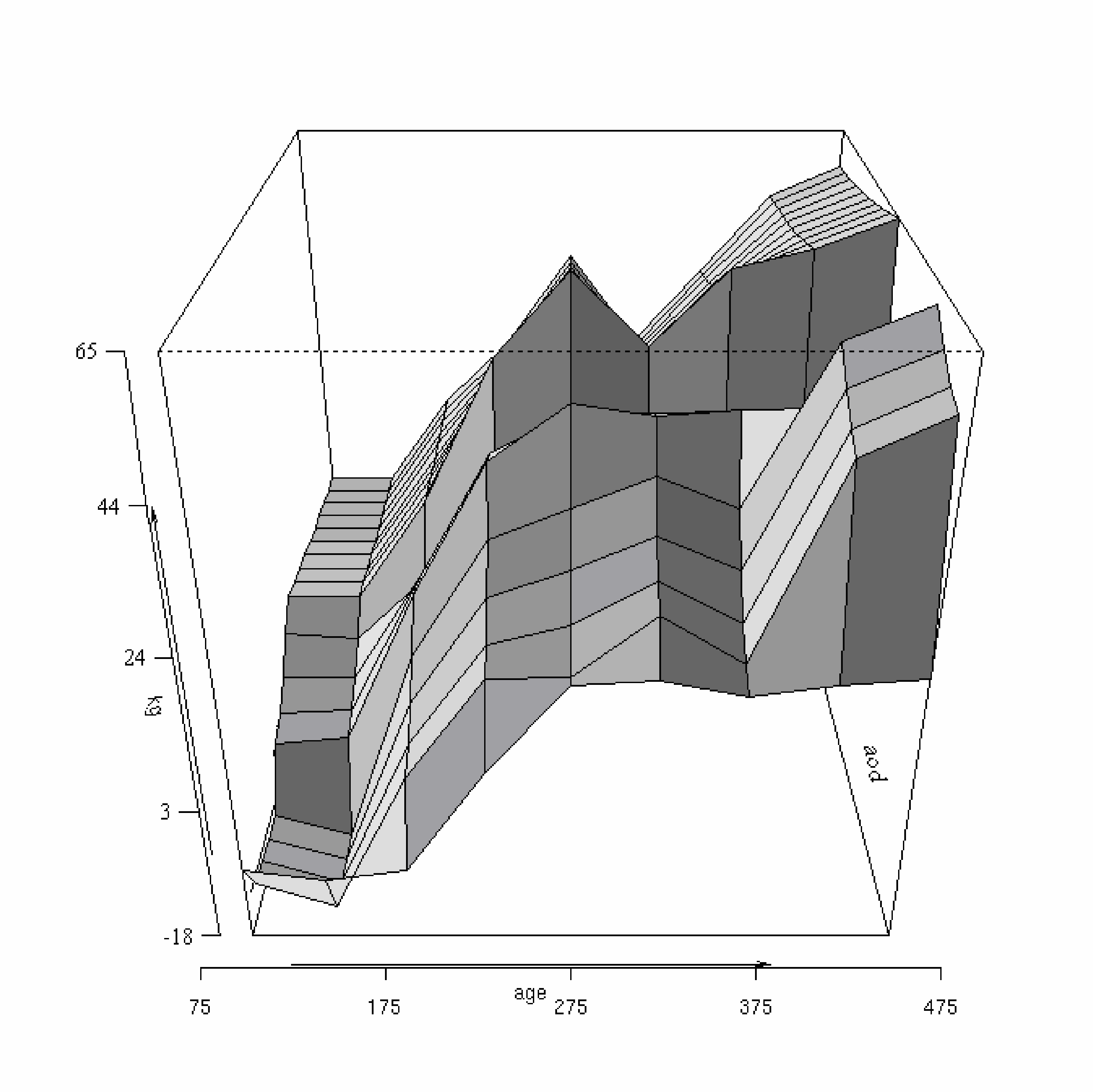
Figure 4.2. Plot of age of dam effect by age[a]



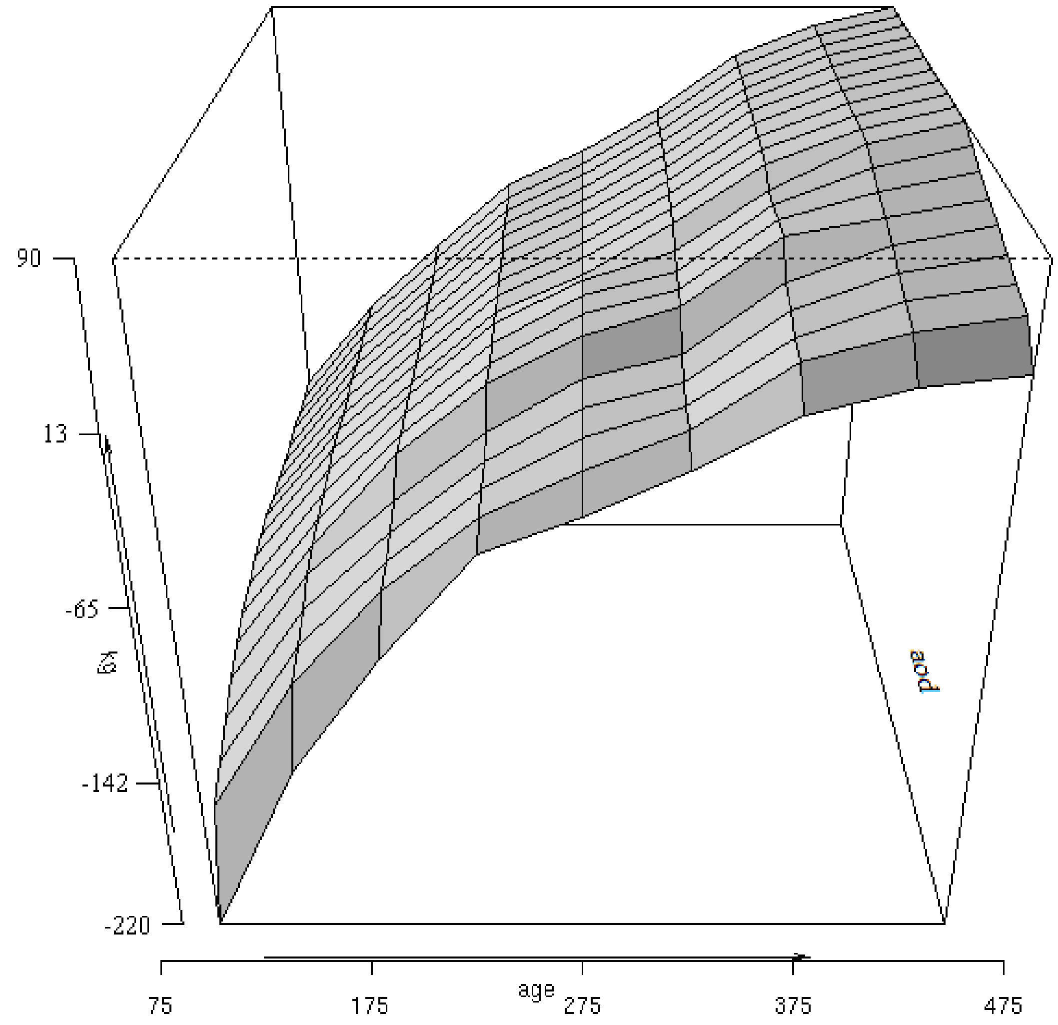
[a] The nested polynomial model is represented by the dashed line. The two-dimensional spline with weighted spline extrapolation is represented by the solid line.

Figure 4.3. 3-D plots of two-dimensional splines

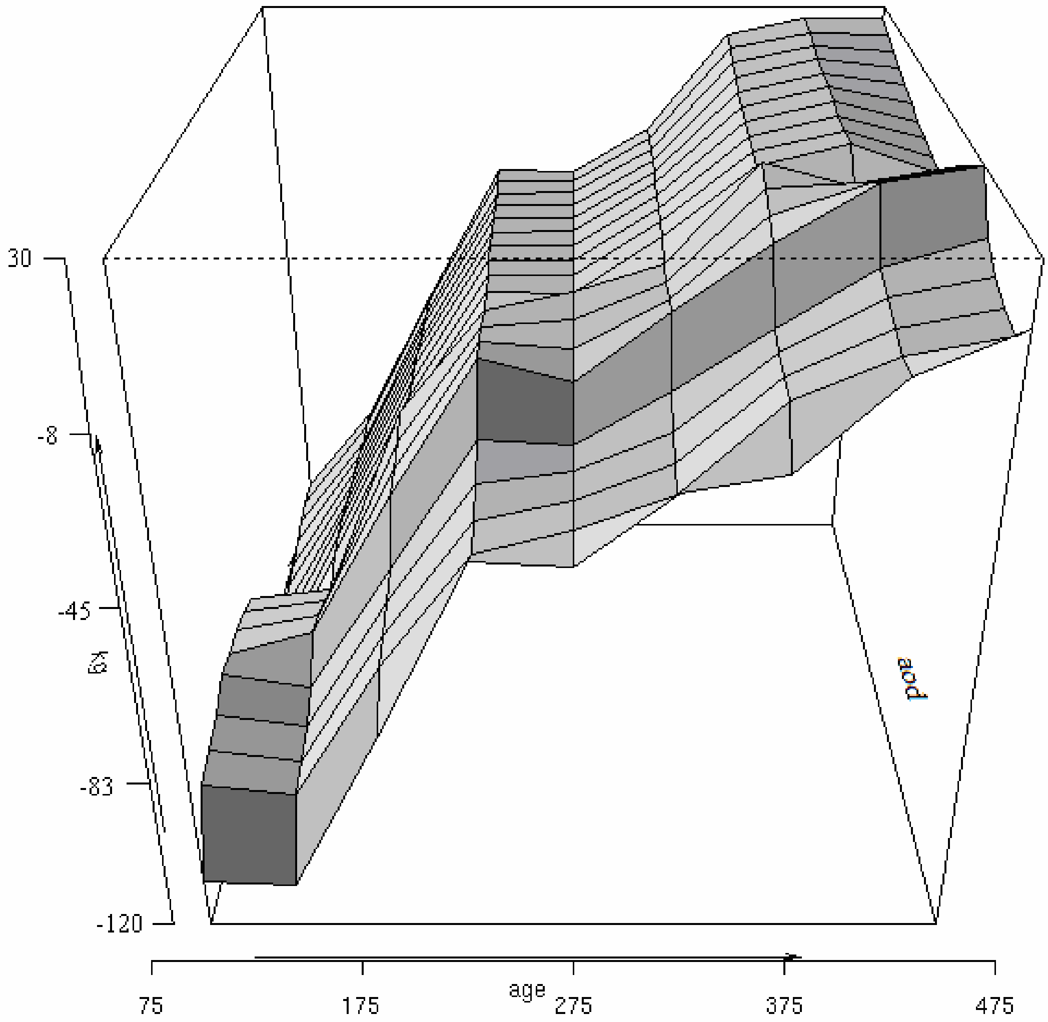
a.



b.



c.



CHAPTER 5

CONCLUSIONS

The results of these studies have shown that computationally feasible and robust longitudinal models can be implemented for nation beef cattle evaluations of growth. By deriving parameters from multiple-trait (co)variance matrices, stable and reliable parameters for orthogonal cubic Legendre polynomials can be obtained. Furthermore, diagonalization of Legendre polynomial parameters greatly decreases computational cost. Random regressions utilizing linear splines are computationally simple and have the capability to use existing multiple-trait parameters. Models utilizing splines and Lengedre polynomial required fewer iterations to converge than did the multiple-trait model.

Testing of random regression models showed that two-dimensional splines can effectively estimate fixed effects. The joint modeling of age of dam and age of animal eliminates the need to nest functions based on the animal's age, and abolishes the need to explicitly model age of dam by age of animal interactions. The use of splines functions can also reduce the effects of sparse data fixed effect estimations. The coupling of two-dimensional splines for fixed effect estimation, and computationally feasible random regressions can yield a practical and robust model for evaluation of growth in beef cattle.

APPENDIX A

PARAMETERS USED FOR RANDOM EFFECTS

Table A.1. Additive genetic effects (direct and maternal) for multiple trait models and random regression models utilizing linear splines

	Direct			Maternal		
Trait[a]	Bwt	Wwt	Ywt	Bwt	Wwt	Ywt
Bwt	7.854	18.626	22.104	-0.939	-3.734	-4.091
Wwt		144.909	163.930	-3.881	-15.433	-16.910
Ywt			290.221	-5.584	-22.205	-24.331
Bwt				2.348	2.103	1.569
Wwt					43.228	41.693
Ywt						47.119

[a] Bwt = birth weight parameter. Wwt = weaning weight parameter. Ywt= yearling weight parameter.

Table A.2.Additive genetic effects (direct and maternal) for random regression models utilizing cubic Legendre polynomials.

	Direct				Maternal			
Item[a]	d1	d2	d3	d4	m1	m2	m3	m4
d1	510.104	0	0	0	-43.762	-0.900	-1.885	3.170
d2		0.643	0	0	0.673	0.014	0.029	-0.049
d3			47.938	0	-5.130	-0.105	-0.221	0.372
d4				69.935	0.954	0.020	0.041	0.069
M1					63.023	0	0	0
M2						10.845	0	0
M3							0.250	0
M4								2.567

[a] d = direct effect parameter for each of the four orders of the cubic Legendre polynomial. m= maternal effect parameter for each of the four orders of the cubic Legendre polynomial.

Table A.3. Maternal permanent environmental effect parameters for multiple trait models and random regressions utilizing linear splines

Trait[a]	Bwt	Wwt	Ywt
Bwt	0.040	0.986	0.976
Wwt		45.600	49.496
Ywt			57.100

[a] Bwt = birth weight parameter. Wwt = weaning weight parameter. Ywt= yearling weight parameter.

Table A.4. Maternal permanent environmental effects for random regressions utilizing cubic Legendre polynomials

Item	mpe1	mpe2	mpe3	mpe4
mpe1	0.002	0	0	0
mpe2		2.413	0	0
mpe3			3.899	
mpe4				78.375

[a]mpe = maternal permanent environmental parameter for each of the four orders of the cubic Legendre polynomials.

Table A.5. Permanent environmental effect parameters for random regression models utilizing linear splines

Trait[a]	Bwt	Wwt	Ywt
Bwt	1.031	7.240	8.998
Wwt		187.511	298.101
Ywt			557.360

[a] Bwt = birth weight parameter. Wwt = weaning weight parameter. Ywt= yearling weight parameter.

Table A.6. Permanent environmental effect parameters for random regression models utilizing cubic Legendre polynomials

Item[a]	pe1	pe2	pe3	pe4
pe1	0.071	0	0	0
pe2		24.033	0	0
pe3			117.598	
pe4				925.211

[a] pe = permanent environmental parameter for each of the four orders of the cubic Legendre polynomials.

Table A.7. Residual parameters for multiple trait models

Trait[a]	Bwt	Wwt	Ywt
Bwt	9.850	9.906	10.160
Wwt		344.700	345.800
Ywt			728.500

[a] Bwt = birth weight parameter. Wwt = weaning weight parameter. Ywt= yearling weight parameter.

Table A.8. Residual parameters for random regressions utilizing linear splines and cubic Legendre polynomials

Trait[a]	Bwt	Wwt	Ywt
Bwt	8.831	0	0
Wwt		121.944	0
Ywt			214.396

[a]Bwt = birth weight parameter. Wwt = weaning weight parameter. Ywt= yearling weight parameter.