Breeding and Genetics: Statistical Methods I

21 Joint modeling of age of dam and age of animal for growth in Gelbvieh by the random regression model. K. Robbins*, I. Misztal, and J. Bertrand, *University of Georgia*, *Athens*.

We examined the joint modeling of age of dam and age of animal in a random regression model analysis of growth in Gelbvieh cattle. The first method (M1) was analogous to multiple trait analysis and consisted of age of dam as a 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 of animal knots at birth, 160d, 205d, 250d, 320d, and 365d; and age of dam knots at 725d, 1464d, and 2189d. A data set containing Gelbvieh growth records was split along contemporary groups (cg) into two data sets. Data set 1 (D1) contained 643,937 records and was used for prediction by mixed model equations. Data set 2 (D2) contained 476,198 records and was used for cross validation. Models were evaluated based on average squared error (ASE) and plots of solutions. ASE for weights associated with birth weight, weaning weight and yearling weight for M1 were 82, 3042, and 6203. For M2 the ASE were 82, 3490, and 6722. Plots of the data showed that weight decreased after 350d. This was traced to a correlation between the age at which animals yearling weights were recorded and the animal's daily gain. This confounding had the greatest effect on the two-dimensional splines. After splitting the splines into two groups, the ASE for M2 at yearling weight declined to 5908. M1 modeling was better in regions with denser data while M2 performed best with sparser data. Two-dimensional modeling of fixed effects is complicated when records are concentrated in narrow age ranges and confounding exists between variables.

Key Words: Polynomial Regression, Two-Dimensional Spline, Cross Validation

22 Analysis of first insemination success subject to uncertainty in Australian Angus cattle. M. L. Spangler*, R. L. Sapp, R. Rekaya, and J. K. Bertrand, *The University of Georgia*, *Athens*.

Field data consisting of 33,099 records from Australian Angus herds were used to investigate two methods of analyzing uncertain binary responses for success or failure at first insemination. A linear mixed model, at the liability scale, that included herd, year, and month of mating as fixed effects; unrelated service sire, additive animal and residual as random effects; and linear and quadratic effects of age at mating as covariates was used to analyze the binary data. An average gestation length (GL) and standard deviation (SD), by sex, derived from artificial insemination (AI) records and observed days to calving (DC) interval were used to assign binary conception data from bull-sired females. Females deviating from average GL leads to uncertain binary responses. Two analyses were carried out: 1) a threshold model fitted to uncertain binary data ignoring uncertainty (M1), and 2) a threshold model fitted to uncertain binary data accounting for uncertainty via fuzzy logic classification (M2). There was practically no difference between point estimates obtained from M1 and M2 for service sire (0.135 and 0.127) and herd variances (0.128 and 0.123). However, the estimates of the additive variance from M1 and M2 were 0.055 and 0.031, and the corresponding heritability estimates were 0.042 and 0.024, respectively. Pearson correlations indicated no major re-ranking would be expected for service sire effects (0.99) and animal breeding values (0.98) using M1 and M2. Given the results of the current study, for noisy binary data, a threshold model contemplating uncertainty is suggested to avoid bias when estimating genetic parameters. The current study is being extended to situations were uncertain binary responses are jointly analyzed with other continuous/binary traits.

Key Words: Beef Cattle, Fuzzy Logic, Fertility

23 Properties of random regression models using linear splines. I. Misztal*, *University of Georgia*, *Athens*.

The purpose of this study was to determine rules to select knots in random regression models using linear splines (RRMS). Such models are much easier to implement than models with polynomials because of superior numerical properties and simplicity of obtaining parameters. Parameters in RRMS are similar to those in multiple trait models (MTM) with traits corresponding to knots, with the exception that the residual variance in MTM equals the sum of residual + permanent environmental variances in RRMS. The variance of an effect with linear splines, relative to a straight line, is concave between adjacent knots. The maximum depression is in the middle of knots and equals 0.5(1-r), where r is the correlation between the knots. For RRMS, the vector of covariables is [0.. 1-t t ..0], where t=<0,1>. The concavity in the middle can be eliminated if this vector is modified to [0.. $(1-t)^q$ t^q ..0], where q=log[2(1+ r)]/[2log(2)]. The influence of correlations between knots on accuracy and variance of predictions was analyzed by simulation. The model to simulate the data included 5 knots spaced equally, with correlations decreasing linearly with increasing knot distance; the correlation between the extreme knots was varied from 0.0 to 0.99. Observations were simulated for trajectory points corresponding to each knot for 1000 unrelated individuals with 50 observations per individual. Predictions were obtained for points corresponding to extremes and to the middle of the trajectory by models with 5 knots (K5), 2 knots (K2), and 2 knots with covariables modified (K2M). Compared to K5 and depending on r, the predictions by K2 (K2M) were inflated at the extremes by 0-16% (0-3%) and deflated in the middle by 0-39% (0-2%). Accuracies by K2M and K2 were similar and 0.0-0.04 below those by K5 but < 0.01 below for r ≥ 0.6. With modified formulas, the accuracy of RRMS increased only marginally when the correlations between the knots were ≥ 0.6. In practical analyses, knots in RRMS can be selected so that 1) they cover the entire trajectory and 2) the correlations between the adjacent knots are ≥ 0.6.

Key Words: Random Regression Model, Linear Splines

24 Calculating the distribution of the correlation between estimated breeding values from different analyses. D. Garrick*, *Colorado State University*, *Fort Collins*.

The validation of EPDs or EBVs from different genetic evaluations is a common occurrence. Two circumstances can be distinguished. 1) Different datasets are used for two analyses as in comparison of sire EPDs from two different progeny tests. 2) Datasets reflect a part-whole relationship as in estimates from a subset of data (eg region) that was also included in a larger (eg national) analysis. When either of the estimates are not perfectly accurate, the correlation between estimates are biased downwards. Accordingly, it is difficult to interpret the observed correlation. However, its expected distribution can be readily simulated. Consider the case with different datasets in the two analyses. Define a vector **u** of true and BLP/BLUP estimated effects on the same non-inbred animal in analyses 1 & 2 as containing g_1, g_2, $g\hat{}_1$, $g\hat{}_2$. Lower triangular elements of var(**u**) are (by row): var(g_1), cov(g_{12}), var(g_2), r_1^2 var(g_1), r_1^2cov(g_{12}), r_1^2 var(g_1), r_2^2cov(g_{12}), r_2^2 var(g_2), $r_1^2r_2^2$cov(g_{12}), r_2^2var(g_2) where r_i is the correlation between true and estimated merit in dataset i. This matrix forms the diagonal blocks of the var-cov matrix of all *n* animals represented in the two datasets. The Cholesky decomposition of such a matrix, in product with a vector of *4n* independent standard normal deviates will produce one possible realization of true and estimated values. The observed correlation (or regression) can be calculated for this one sample. This procedure can be repeated *k* times (say *k*=1000) to get a distribution of the expected correlation (or regression). A one-sided critical value at say α=0.05 can be obtained by sorting the sample correlations and using the 50th (from lowest) value. A similar procedure can be used for part-whole datasets. In that case, the expected covariance between estimates reduces from $r_1^2r_2^2$cov(g_{12}) to r_1^2cov(g_{12}) where dataset 1 is the subset of dataset

2. Inbreeding and relationships can be accounted for during simulation by including the Cholesky decomposition of the numerator relationship matrix in the calculation.

Key Words: Genetic Evaluation, Selection Index

25 A bivariate quantitative genetic model for a linear Gaussian trait and a survival trait. L. H. Damgaard* and I. R. Korsgaard, *Research Centre Foulum, Dept. Genetics and Biotechnology, Bioinformatics and Statistical Genetics*, *Tjele, Denmark*.

A bivariate quantitative genetic model for a linear Gaussian and a survival trait genetically and environmentally correlated was derived and implemented. For the survival trait, we considered the Weibull log-normal animal frailty model. A Bayesian approach using Gibbs sampling was adopted. Model parameters were inferred from their marginal posterior distributions. The required fully conditional posterior distributions were derived and issues on implementation discussed. The two Weibull baseline parameters were updated jointly using a Metropolis-Hasting step. The remaining model parameters with non-normalized fully conditional distributions were updated univariately using adaptive rejection sampling. Simulation results showed that the estimated marginal posterior distributions covered well and placed high density to the true parameter values used in the simulation of data. All the true parameter values were within the 95% central posterior density regions defined by the 2.5% and 97.5% percentiles. In conclusion, the proposed method allows inferring additive genetic and environmental correlations between a linear Gaussian trait and a survival trait.

Key Words: Survival, Gaussian, Genetic Correlation

26 Bivariate recursive and simultaneous models for milk yield and somatic cell scores. G. de los Campos*[1], D. Gianola[1], and B. Heringstad[2], [1]*University of Wisconsin-Madison*, *Madison*, [2]*Norwegian University of Life Sciences*, *Aas, Norway*.

Diseases may affect production and vice versa. Standard linear model theory does not accommodate recursiveness or simultaneity of effects. Structural Equation Models (SEM), however, allow modeling such features. Using LISREL®8, we compared four bivariate SEM for analysis of milk yield (MY) and somatic cell scores (SCS). Models were: MO (standard), M1 (SCS=>MY), M2 (SCS<=MY), and M3 (SCS<=>MY); arrows indicate direction of effects. The data set had test-day MY and SCS, and clinical mastitis (CM) records of 33,453 first-lactation daughters of the 245 Norwegian Red (NRF) sires with a first progeny test in 1991 or 1992. First lactation was divided into five 60-day periods and a test-day was assigned to each period. Within-herd SCS and MY deviates were responses, and presence of CM within 15-days prior to test day, age at calving, and sire were 'exogenous' variables. The Bayesian Information Criterion (BIC) favored M1. SCS had a negative effect on MY both in M1 and M3 (in M1: -1.1 kg/day/SCS, $p < .001$). The association between SCS and MY was mostly due to a negative effect of SCS on MY; 'dilution' effect (MY=>SCS) seems unlikely to exist. Using estimates from M1, an event of CM would be expected to increase SCC by 70,000 cells/ml in the following test day; through the recursive effect (SCS=>MY), MY would be reduced by 0.93 kg/day. Estimates may be biased downwards, because of false negative CM (cases outside of the 15-day period may affect SCS). For M1, phenotypic (additive genetic) variances of MY and SCS were 13.19 (1.74) and 1.19 (0.11) respectively. Phenotypic and genetic correlations between SCS and MY were -0.23 and 0.34 respectively. The phenotypic correlation was the most sensitive parameter to specification of recursive effects.

Key Words: Diseases and Production, Structural Equation Model, Simultaneity

27 Standard errors of solutions in large scale mixed models, application to linear and curvilinear effects of inbreeding on production traits. N. Gengler*[1,2] and C. Croquet[1,2], [1]*National Fund for Scientific Research*, *Brussels, Belgium*, [2]*Gembloux Agriculture University*, *Gembloux, Belgium*.

Many approaches for using linear mixed models do not produce standard errors of solutions. However, knowing the standard errors allows for statistical tests. Even if exact estimation of standard errors is not feasible in large mixed models, there are methods to approximate them. We based this on Mixed Model Conjugate Normal Equations associated with a Preconditioned Conjugate Gradient (PCG) solver. The advantage of associating both methods is that the right hand side vector normally accumulated by PCG can be easily changed to a function of solutions vector $\mathbf{k}$ allowing direct solution for $\mathbf{\Phi}=\mathbf{C}^{-1}\mathbf{k}$ using regular PCG solving programs. The square root of $\mathbf{k}'\mathbf{\Phi}=\mathbf{k}'\mathbf{C}^{-1}\mathbf{k}$ gives the standard error associated with the function of solutions described by $\mathbf{k}$. Often a block of $\mathbf{C}^{-1}$ is needed. Its elements were obtained by computing linear functions of element of this block and by backsolving to obtain the needed elements. In matrix notation let $\mathbf{K}$ be the coefficients of the linear functions and $\mathbf{D}$ a matrix containing the values obtained by computing $\mathbf{K}'\mathbf{C}^{-1}\mathbf{K}$. The elements of the block were then obtained as $(\mathbf{KK}')^{-1}\mathbf{KDK}'(\mathbf{KK}')^{-1}$. This method was applied to study linearity of inbreeding depression on milk, fat and protein test-day yields. Inbreeding effects were estimated using linear, quadratic and cubic regressions on inbreeding coefficients inside breeds in a test-day model similar to the one used in the Walloon Region of Belgium. The pedigree contained 956,516 animals. A total of 5,596,038 first lactations test-day records from 660,407 cows were used. Results had contrasting behaviors, however evaluation of plotted inbreeding effect and the associated confidence interval showed that between 0 an 10% inbreeding differences among evaluations of inbreeding depression were small.

Acknowledgements: Coraline Croquet who is Research Fellow, and Nicolas Gengler, who is Research Associate of the National Fund for Scientific Research, acknowledge their support.

Key Words: Standard Error, Inbreeding, Curvilinear

28 Predictions of test day yields for milk production traits in cattle by partial least squares multiple regression. N. P. P. Macciotta*[1], D. Vicario[2], C. Dimauro[1], N. Bacciu[1], and A. Cappio-Borlino[1], [1]*Università di Sassari*, *Sassari, Italia*, [2]*Italian Association of Simmental Cow breeders*, *Udine, Italia*.

The research of methods able to predict Test Day (TD) yields from a limited number of actual records available is an important challenge for the dairy cattle industry. Most of proposed methods deal with a univariate approach and can forecast only future TD in lactation in progress. On the other hand, multivariate approaches are theoretically and computationally heavy. The Partial Least Squares Regression (PLS) multivariate approach can represent a valid alternative, being able to handle plans characterised by the presence of missing data in different parts of the lactation. Moreover, the extraction of orthogonal latent factors enable the PLS to reduce problems of the collinearity among predictors and, at the same time, to exploit correlations between dependent and independent variables. The PLS prediction ability was tested on a data set of 31,356 lactations of Italian Simmental Cows of parity 1 to 3, with 8 test day records of milk production traits (milk, fat and protein yields) per lactation, arranged in a multivariate setting. Ten scenarios of missing TD records were simulated. Predictions were calculated separately for each parity class. Correlations among actual and predicted TD yields evaluated by cross validation methodology ranged from 0.60 to 0.90 for protein and from 0.55 to 0.80 for fat in the scenarios where also milk yield has to be predicted. Correlations increased up to 0.97 for protein and 0.88 for fat when all milk TD were available. Average correlations between actual and predicted TD for milk yield was 0.87. The analysis of Mean Square Error of Prediction confirm the higher accuracy of the PLS method and highlights a certain degree of imprecision mainly due to the random nature of individual variation. Results of the study indicate a good predictive ability of the PLS method that, in addition, is markedly flexible, does not require special computing capability and is easily transferable to the farm level.

Key Words: PLS, Prediction, Milk Test Day

29 Genetic parameters of latent variables related to main traits of lactation curve shape. N. P. P. Macciotta*[1], D. Vicario[2], and A. Cappio-Borlino[1], [1]*Università di Sassari*, *Sassari, Italia*, [2]*Italian Association of Simmental Cow Breeders*, *Udine, Italia*.

The genetic analysis of multivariate phenotypes has to cope with computational issues essentially due to the large number of parameters to be estimated. In recent years, dimension-reduction techniques have been proposed mainly based on principal component analysis. However, this multivariate technique does not always allow a simple and meaningful interpretation of the leading new variables. A different technique to extract latent variables able to reconstruct the (co)variance matrix of original data is the multivariate factor analysis, where extracted factors can be rotated to make their interpretation easier. In this work the factor analysis was used to extract two latent variables from Test Day milk yields, taken at different days in milk, related to peak yield and lactation persistency. A data set of 48374 lactations of 21721 Italian Simmental cows was used. Edits were on the number of Test Day records for cow (8), parity (<7), days in milk at first test (<30), lactation length (>200), n. cows per herd (>5). TD milk yields were arranged in a multivariate setting, latent factors were extracted. Factor scores were then used as new variables and analysed by a bivariate animal model that included the fixed effects of the herd, parity, year and calving season, and the random effects of the genetic additive value of the animal and of the permanent environment. Estimates of heritability are moderately low, (0.14 both for persistency and peak yield) confirming the results reported from other authors. Repeatabilities show values of 0.27 for peak and 0.29 for persistency. Particularly interesting is the moderate genetic correlation among these two variables (0.26) suggesting the possibility to select, at least to a certain extent, separately on different aspects of lactation curve shape.

Key Words: Lactation curve shape, Persistency of lactation, Latent factors

30 Simultaneous estimation of environmental values and genetic parameters in reaction norm model. G. Su*, P. Madsen, M. S. Lund, D. Sorensen, I. R. Korsgaard, and J Jensen, *Danish Institute of Agricultural Sciences, Department of Genetics and Biotechnology*, *Tjele, Denmark*.

The reaction norm model is becoming a popular approach for the analysis of G x E interactions. In a classical reaction norm model, the expression of a genotype in different environments is described as a linear function (a reaction norm) of an environmental gradient, such as overall effects of various environments (environmental values). An environmental value could be defined as the mean performance of all genotypes in a particular environment, which is typically unknown. One approximation is to estimate the mean performance based on data then treat this as a known covariate in the model (referred to as the approximate method below). However, a more satisfactory alternative is to infer the environmental values simultaneously with the other parameters of the model. The objectives of this study were (1) to describe a method and its Bayesian MCMC implementation that makes this possible and (2) to confirm the superiority of the present method relative to the approximate method by a simulation study. In the simulation study, data were generated using simulated herd-year effects as covariates of the reaction norm. The correlation between the estimates of herd-year effects from the present method and the true values was close to one. The genetic parameters inferred from the present method were similar to those estimated from a reaction norm model using true values of herd-year effects. On the other hand, using phenotypic mean of herd-year as a proxy for the environmental value resulted in biased estimates of genetic parameters.

Key Words: Reaction Norm, G x E Interaction, Genetic Parameters

Dairy Foods: Dairy Chemistry

31 Influence of lipolysis and proteolysis of calibration milks on infrared milk analyzer performance. K. E. Kaylegian* and D. M. Barbano, *Cornell University*, *Ithaca, NY*.

Lipolysis was measured weekly as the increase in free fatty acid (FFA, meq/kg milk) content and proteolysis was measured weekly as the decrease in casein as a percentage of true protein (CN%TP). Pasteurized, potassium dichromate preserved modified milk (MM) calibration sets had a shelf life of 4 wk and raw preserved producer milk (PM) calibration sets had a shelf life of 2 wk. Every day that the chemical analyses for lipolysis and proteolysis were performed, MM and PM calibration sets were used separately to calibrate a fixed-filter mid-infrared (MIR) analyzer. The experiment was replicated 3 times. The MM sets had a smaller ($P < 0.01$) mean and change in lipolysis over the first 2 wk of use than the PM sets. The mean FFA levels were 0.115 and 0.253 meq/kg milk for the MM and PM sets, respectively. The mean decrease in CN%TP was larger ($P < 0.01$) for the MM sets (1.96%) at the end of the first 2 wk than for the PM sets (1.54%). The mean level of proteolysis observed for the MM sets at the end of the 4 wk set life was a 4.06% decrease in CN%TP, and no effects on the calibration slope and intercept were observed . The highest FFA level observed was 0.589 meq/kg milk for an individual PM sample at the end of the 2 wk set life. Individual PM samples with high FFA levels did show a larger difference between IR predicted value and chemistry for those samples, as expected, but these did not affect the protein slope and intercept because they were in the middle of the protein concentration range of the calibration set. No significant differences were observed in the MIR calibration slope and intercept that were attributed to lipolysis or proteolysis for either type of calibration set.

Key Words: IR Milk analyzer Calibration, Lipolysis, Proteolysis

32 Comparing a gas chromatography/mass spectrometry technique with sensory evaluation in relation to the acceptability of fluid milk. A. A. Glueck-Chaloupka*[1], C. H. White[2], and W. E. Holmes[3], [1]*The Kroger Company*, *Cincinnati, OH*, [2]*Mississippi Agricultural & Forestry Experiment Station*, *Mississippi State, MS*, [3]*Mississippi State Chemical Lab*, *Mississippi State, MS*.

A gas chromatography/mass spectrometry (GC/MS) technique was evaluated for ability to detect changes in volatile compounds in reduced-fat milk over time. Pasteurized reduced-fat fluid milk samples were collected from 10 filler heads on a fluid milk packaging machine from one dairy plant (n=150). Expert sensory panelists evaluated samples for acceptability using the ADSA scorecard and a quality scale. Ten volatile compounds known to have an impact on the flavor and flavor intensities of reduced-fat fluid milk over the shelf life of the product were monitored with a GC/MS technique. The GC/MS technique was able to provide identification and quantification of compounds. It provides a qualitative and quantitative means of identifying metabolites present in both "good" and "bad" quality milk. Both sensory and GC/MS evaluations occurred on days 0, 1, 7, 11 and 15 at 7°C. Sensory data were subjected to analysis with Statistical Analysis Software (SAS) version 8.0. Response factors were determined by dividing the concentrations by the area counts and reported for the volatile compounds. By day 7 expert panelists were able to detect overall flavor intensity differences ($P<0.05$). Overall flavor scores and occurrence of "pleasant" flavor attributes decreased over shelf life. Concentrations on day 0 versus day 15 of ethanol (2.11ng, 2.50ng), ethyl butyrate (0.00ng, 0.01ng), ethyl hexanoate (0.01ng, 0.22ng), 3-methyl-1-butanol (0.00ng, 0.01ng), 2-nonanone (0.02ng, 0.12ng), and 2-heptanone (0.002ng, 0.014ng) increased as the milk aged. A similar trend over the 15 day evaluation period for both testing methods was shown, indicating that as the milk ages, concentration of volatile compounds increase, and overall flavor scores and occurrence of "pleasant" flavor attributes decrease. Therefore, GC/MS, when combined with sensory evaluation, shows promise in understanding flavor deterioration in reduced-fat milk.

Key Words: Reduced-fat milk, GC/MS, Flavor

33 Novel technique for the differentiation of caseins and whey proteins in confocal scanning laser microscopy. A. Dubert-Ferrandon*, A. Grandison, and K. Niranjan, *The University of Reading*, *Whiteknights, Reading, UK*.