NANP Nutrition Models Workshop

1 Welcoming remarks. J. McNamara*, McNamara Research in Agriculture Firm, Pullman, WA.

The animal sciences are made up of a wide range of processes from single gene inheritance through cellular, organ, whole-animal, herd, farm, system, and economic and social processes. Research over the last 100 years has given us tremendously detailed insight and massive data sets on all aspects of food animal production. Nutrition and feeding, including all the interactions with genetics, the environment and reproduction are central to supplying health and efficient animals for a healthy human population. Although no one or one research group can simultaneously teach or do research in all these sections, our charge to improve the security, safety and efficiency of the animal food product system dictates that several efforts integrate 2 or even many of these systems toward a common end goal. Many government, NGO and private companies and organizations now are involved in and in fact require complex approaches to solving the complex issues of food production. Various research and educational approaches are valid and useful, but they all make use of sound quantitative biological and statistical approaches, including various types of mathematical and statistical modeling techniques. The National Animal Nutrition Program and the National Research Support Program exist to support and encourage excellence in education and research in Animal Sciences Nutrition and all aspects of the food system that are integrated with that (we don’t feed animals in a separate system, the inputs into the animals and outputs from them are an integral part of the system as a whole). The workshop today encompasses a variety of useful concepts and practices that will allow the student or scientist at any stage of their career to more fully understand and use a variety of biological, mathematical, and statistical approaches to systems biology research and education. The topics, practices, and information will be useful for anyone starting a career in the animal sciences or wishing to integrate their existing teaching and research into a relevant systems approach to providing a wholesome food supply.

Key Words: systems biology, nutrition, modelling

2 Tutorial on R software. T. Hackmann*, University of California, Davis, Davis, CA.

R statistical software will be used throughout the NANP Nutrition Models Workshop. The objective of this tutorial is to familiarize workshop attendees with use of this software. Specifically, this tutorial will walk attendees through basic data manipulation, visualization, and other functions. Assistance will be available for attendees who have difficulty installing the software on their laptops.

Key Words: R, tutorial, modeling

3 Estimation of parameter values: Lecture and exercises. M. D. Hanigan*1 and V. L. Daley1,2, Virginia Tech, Blacksburg, VA, National Animal Nutrition Program, University of Kentucky, Lexington, KY.

Application of nutritional knowledge generally requires expression in mathematical form. It is insufficient to conclude that animals should be fed more of a nutrient; the recommendation must include an estimate of how much more. Thus, models that accurately and precisely represent animal responses to varying nutrient supply are a critical product of nutrition science. They also allow quantitative hypothesis testing which guides the scientific process. Hence, the construction of models and the derivation of parameter estimates for those models are a critical component of nutrition science. For this learning exercise, it is assumed the participant is proficient in the use of R and the use of linear and nonlinear regression functions, and has participated in the National Animal Nutrition Program Level 1 Workshop, which includes a module on building a portion of this model, or has gained that expertise through self-study. Participants will be given a data set and a model containing 4 components that interact. These are (1) a pool of insoluble N in the rumen, (2) a pool of soluble N in the rumen, (3) a pool of ruminal microbes that are consuming soluble N, and (4) digestion and absorption of N from the small intestine. The model will be fitted to observed ruminal and fecal N outflow data to derive parameter estimates for the conversion of insoluble N to soluble N and the fractional use of the soluble N in support of microbial growth, and digestion of N in the intestines. The exercise demonstrates the use of an optimizer to fit model parameters to observed data.

Key Words: mathematical model, parameter estimation, instruction

4 Cross validation and bootstrapping: Part 1 (lecture). J. A. D. R. N. Appuhamy*1 and L. E. Moraes2, Department of Animal Science, Iowa State University, Ames, IA, Department of Animal Science, The Ohio State University, Columbus, OH.

Cross-validation (CV) and bootstrapping are resampling methods that refit a model to samples drawn from the data. CV helps in choosing a “best” model associated with the lowest prediction error rate, whereas bootstrapping allows determining the uncertainty of parameter estimates. One may be tempted to use whole data set to develop and evaluate a model simultaneously. This approach however has issues such as overfitting and thus selects models that would potentially fail on an independent data set. These limitations can be overcome successfully with CV. Traditionally, CV is applied by splitting the data into 2 sets training, test that are used for model development and evaluation, respectively. This method called Hold-out is not recommended particularly for small data sets as the error rate would depend exclusively on the split and be misleading for a different split. Data splitting methods such as K-fold and Leave-one-out are recommended to overcome those limitations. K-fold CV involves dividing the data into K number of samples and holding out one as the test set to determine the error rate. In leave-one-out CV, only one observation is held out at a time as the test set. In both cases, the true error rate for models with continuous responses is generally estimated as the average of the separate error estimates. Bootstrapping is a powerful statistical tool involving resampling with replacement and commonly used to quantify standard error or the confidence interval of statistical estimates. Consequently, bootstrapping allows for determining bias, standard errors, and confidence intervals of statistical estimates. Traditionally, the uncertainty of model parameters are estimated by deriving the sampling distribution based on assumptions about distribution of the population. In contrast, bootstrapping allow estimating the uncertainty without explicitly deriving the sampling distribution that way although it is important to keep in mind that the bootstrap depends on the bootstrap principle “Sampling with replacement behaves on the original sample the way the original sample behaves on a population.”
This lesson will cover the principles and implementations of CV and bootstrapping for models frequently used in animal nutrition.

Key Words: model evaluation, prediction error, resampling

5 Cross validation and bootstrapping: Part II (exercises). J. A. D. R. N. Appuhamy*1 and L. E. Moraes2, 1Department of Animal Science, Iowa State University, Ames, IA, 2Department of Animal Science, The Ohio State University, Columbus, OH.

Here we demonstrate a few applications of cross validation and bootstrapping in evaluating the predictive ability and determining uncertainty of the parameter estimates of a linear regression model using R, a freely available and widely used statistical programming language. The packages such as “design,” “DAAG,” “caret,” and “boot” are capable of performing cross validation of linear models in R. The “bootstrap” package particularly provides extensive facilities for bootstrapping and thus estimating the standard error or confidence interval of a single statistic (e.g., mean), or a vector (e.g., regression coefficients). A data set including a given number of enteric methane emission (CH4) measurements, and corresponding dry matter intake (DMI) and dietary fat content is used. A simple linear regression model to predict CH4 is first developed including DMI and evaluated separately using Hold-up, K-fold, and Leave-one-out cross validation methods. The outputs are discussed and the methods are compared related to the variability of MSPE, and computational cost. The K-fold cross validation is performed with traditional K = 10 (90% of data for training and 10% for test), and compared with lower (K = 5) and higher (K = 20) number of folds. One of the issues with K-fold cross validation is that it often has a high variability, if performed multiple times on the same data. The replicated K-fold cross-validation method addresses this issue by performing the whole process several times averaging over replications. Therefore, we perform replicated K-fold cross validation and compare the MSPE with previous values.

We then use our simple prediction model to demonstrate an application of nonparametric bootstrapping to estimate bias, standard error, and 95% confidence interval of the parameter estimate. Histograms and normal quantile-comparison plots for the bootstrap replications are obtained and discussed. If time permits, the bootstrapping will be repeated with a multiple regression model including both DMI and dietary fat content. The data and all the R scripts will be available in advance for download.

Key Words: confidence interval, K-fold cross validation, standard error

6 Automated model selection: Part I (lecture). V. L. Daley*1, T. J. Hackmann2, and M. D. Hanigan1, 1National Animal Nutrition Program (NANP), University of Kentucky, Lexington, KY, 2University of California, Davis, CA, 3Virginia Tech, Blacksburg, VA.

Automated model selection (AMS) is a procedure to select the best model from a set of candidate models (multi-model inference). This approach can be very useful when the investigator is dealing with a large number of predictor variables to explain a subject of interest (dependent variable). The objective of this lecture is to present the concepts of AMS and illustrate how this procedure can be used in the development of empirical models in Animal Science. The attendees should have some experience in data analysis and empirical models. At the beginning, the hypothesis, objectives, and potential variables associated with the subject of study will be discussed. Then, key concepts in development of a meta-analytic data set and examples from the NANP website (https://animalnutrition.org) will be presented. Data should be assessed for biological coherence and outliers need to be removed. For AMS, a mixed model is fitted using all predictor variables that potentially affect the dependent variable (global model), then a set of sub-models are derived from the fixed terms of the global model. The parameters of those models are estimated by the maximum likelihood method. The AMS approach can use one or more information criterion, but the Akaike’s information criterion corrected for small sample size (AICc) is often adopted. All candidate models are ranked based on the lowest AICc to obtain model weights, then the best set of candidate models are selected. The variance inflation factor (VIF) is used to evaluate the correlation between predictors in the model, and biological coherency of the best models are evaluated. The final stage compares the best candidate models selected (often <10 models) using the root mean square error (RMSE) and the concordance correlation coefficient (CCC). AMS is useful for the development of prediction models when a large set of potential predictor variables are available.

Key Words: multi-model inference, mixed models, Akaike information criterion

7 Automated model selection: Part II (exercises). V. L. Daley*1, T. J. Hackmann2, and M. D. Hanigan1, 1National Animal Nutrition Program (NANP), University of Kentucky, Lexington, KY, 2University of California, Davis, CA, 3Virginia Tech, Blacksburg, VA.

Automated model selection (AMS) can be applied in different research areas for the selection of the best fitting models. The objective of this exercise is to apply the AMS approach as a tool for the selection of prediction models. A hypothetical example will be used to help the audience better understand and apply AMS. This exercise will use RStudio program, which can be freely downloaded from the internet. As an example, the investigator will develop empirical models to predict the dry matter intake of lactating dairy cows using AMS and parallel computation in R (MuMIn). A meta-analytical data set from the National Animal Nutrition Program (NANP, https://animalnutrition.org) is available for this exercise. First, data quality and range will be checked to identify and remove outliers, and data will be visualized (ggplot2). A global mixed model (lme4) will be fitted to the data using all potential predictor variables from the data set. A set of candidate models will be generated using combinations of the fixed terms of the global model. Then, Akaike’s information criterion corrected for small sample size (AICc) is calculated and used to rank the models. The candidate models are collected in a data set named “all models” for future use. The models with the lowest AICc values are collected in another data set named “best models.” The estimated parameters of best candidate models are automatically collected. Only candidate models with a variance inflation factor (VIF) less than 10 are kept in the “best models” data set. Evaluation of the biological coherence and ANOVA to compare the models are performed. The best candidate models will also be evaluated using the root mean squared error of prediction (RMSE) and concordance correlation coefficient (CCC). The best 3 models are automatically collected in a table (sjPlot). During the exercise, additional instructors will be available to help the participants. The attendees will be able to undertake AMS to select the best models and apply those models in research.

Key Words: multi-model inference, empirical models, review

8 Molly and other dynamic models: Part I (lecture). H. A. Rossow*, University of California, Davis, Davis, CA.

Mathematical models are tools to examine existing theories, find gaps in knowledge and explain phenomena such as nutrient digestion and
metabolism. Simulated data from a model can be used to examine model behavior, as in sensitivity analysis, and compared with experimental data to determine if the model makes biological ‘sense’. The objective of this session is to explore how concepts or theories of nutrient digestion, metabolism, and lactation physiology are translated into mechanistic mathematical equations and combined into a whole animal model using the Molly model as an example. Molly is a mechanistic model of a dairy cow composed of a digestive element and an animal element. The digestive element converts chemical composition of the diet to volatile fatty acids, microbial growth and absorbed nutrients using physical attributes of the diet such as proportions of large and small particles and water passage. The animal element converts products from the digestive element into tissues (protein), waste products, heat production or secreted products (e.g., milk, milk fat). Three processes that have been translated into equations in the Molly model will be examined in this session: (1) Passage of digesta through the rumen described in Baldwin et al. (1977; Agric. Syst. 2:255–288); (2) milk production by the mammary gland described in Neal and Thornley (1983; J. Agric. Sci. Camb. 101:389–400); and (3) growth (protein synthesis) described in Oltjen et al. (1986; J. Anim. Sci. 62:86–97). In the lecture, representations of these processes will be examined using conceptual diagrams and differential equations. Then a full lactation simulation will be used to demonstrate how equations of passage, protein accretion, and milk synthesis fit into the model.

**Key Words:** computer simulation model, dairy cow, metabolism

9  **Molly and other dynamic models: Part II (exercises).** H. Rossow*, University of California, Davis, Davis, CA.

Mechanistic models representing physiological processes can be connected to represent whole animal systems. When the models use Michealis-Menton type kinetic equations, examples of ways to connect physiological processes are (1) including central pools (blood pool) of nutrients available for physiological processes, (2) using differential equations that represent nutrient input and output relationships for nutrient pools, and (3) assuming a physiologic process is the same across all tissues, i.e., protein synthesis. Based on the 3 example processes discussed in lecture, (1) passage of digesta through the rumen described in Baldwin et al. (1977; Agric. Syst. 2:255–288); (2) milk production by the mammary gland described in Neal and Thornley (1983; J. Agric. Sci. Camb. 101:389–400); and (3) growth (protein synthesis) described in Oltjen et al. (1986; J. Anim. Sci. 62:86–97), participants will conduct a simulation exercise that uses these modeled processes in Molly to predict lactation performance. In the exercise, participants will observe effects of altering milk production processes on production of the dairy cow to understand how metabolic processes can be represented by mathematical equations to provide a conceptual framework that improves our understanding of animal biology.

**Key Words:** computer simulation model, dairy cow, metabolism