BAYESIAN ANALYSIS IN THE PRESENCE OF MISSING EXPLANATORY VARIABLES: APPLICATION TO DAYS ON FEED IN FEEDLOT CATTLE

By

EL Hamidi A. Hay

(Under the direction of Romdhane Rekaya)

ABSTRACT

Days on feed (DOF) is an essential component in the profitability of feedlot industry. Several equations have been developed to predict DOF that include multiple animal and environmental factors. More recently, molecular information (molecular breeding values) for several traits is available to add to the calculation of DOF, with the advantage that this information is readily available as soon as the animal is genotyped for a set of single nucleotide polymorphisms (SNPs). In this study, a data set of 10,209 feedlot animals were used to establish a prediction equation for DOF based on arrival weight, sex, growth, genotypes (for 200 SNP markers), and an imputed Zilmax status. Three models were compared in their prediction ability of DOF. The first model (M1) included the following effects (sex, HCW, initial weight, marbling score, backfat, ribeye area, and 36 SNPs selected out of the available 200 SNPs). The second model, M2, included all the effects in M1 plus Zilmax status (case/control) based on the predicted probabilities using 0.5 as a threshold value. The third model, M3, included all effects in M1 and the uncertain or unknown Zilmax status. Several analyses were conducted where a joint crisp and soft classification of Zilmax status was adopted.

Keywords: Zilmax, Prediction, Days on feed, SNP

BAYESIAN ANALYSIS IN THE PRESENCE OF MISSING EXPLANATORY VARIABLES: APPLICATION TO DAYS ON FEED IN FEEDLOT CATTLE

By

El Hamidi A. Hay

BS, Southern Polytechnic State University, 2008

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

2011

BAYESIAN ANALYSIS IN THE PRESENCE OF MISSING EXPLANATORY VARIABLES: APPLICATION TO DAYS ON FEED IN FEEDLOT CATTLE

by

El Hamidi A. Hay

Major Professor: Romdhane Rekaya

Committee: Ignacy Misztal

J. Keith Bertrand

Brent Woodward

Electronic Version Approved:

Maureen Grasso

Dean of the Graduate School

The University of Georgia

August 2011

DEDICATION

To: My Family.

ACKNOWLEDGEMENTS

I would like to thank Dr Rekaya for his guidance, support and input. Without him I could not have done it. I would like to also thank the entire animal and breeding genetics group at The University of Georgia.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Days on feed (DOF) is a crucial component in feedlot operations and profit. Accurate prediction of DOF could help feedlot managers in management and marketing decisions. The goal of this study is to develop a statistical model to accurately predict DOF by including management, environmental and genetic factors. Including molecular information is essential especially now with the availability of SNP panels at reasonable costs. A 200 SNP panel was used in this study; this panel was generated from beef candidate gene research, meaning SNPs that are in or near QTLs or genes that have been reported to affect traits in cattle. However, information on some of these factors is not always available in all animals in the feedlot. Such factors include health status, diet, or hormonal treatment. In this study, a data set of 10,209 feedlot animals were utilized to first infer dietary additive (Zilmax) status of feedlot animals missing this information and second to establish a prediction of DOF based on arrival weight, sex, growth, and molecular information consisting of 200 SNP markers. Only 2,449 animals have information about Zilmax treatment; in order to predict Zilmax status a binary response model (case vs control) was adapted and a five-fold cross-validation study was carried out to test the prediction ability. To predict DOF, a statistical model was implemented by including several factors, moreover this model accounts for the potential misclassification of the predicted Zilmax status in the calculation of DOF.

CHAPTER 2

LITERATURE REVIEW

The US beef industry is constantly growing and changing in order to adapt to the increasing consumer demand. The beef industry is mainly divided into three key segments: Cow-calf, stocker feeder, and packer. Feedlots have a special feeding system where cattle are fed high grain diets for a certain number of days to efficiently gain weight and to satisfy carcass and meat quality requirements; this feeding period is referred to as days on feed (DOF) and usually ranges from approximately 100 to 250 days. Several factors can affect DOF and feedlot operations; such factors are management, environment, and genetics. The animal breeding and genetics field utilizes quantitative methods and tools that could help feedlot managers better assess and predict the performance of feedlot cattle. In this study, we developed a statistical model that predicts DOF of feedlot cattle animals based on their phenotypic information and most importantly genetic information.

Animal improvement: from classical animal models to genomic selection

Classical Animal breeding and genetics

Jay Lush is considered the father and pioneer of the animal breeding and genetics field; he developed some of the basic tools and methods used in animal breeding. Dr Lush combined information and statistical tools from Drs. R.A. Fisher and Wright to solve animal breeding problems. For example, he explained how selection can result in inbreeding and how inbreeding can affect diversity among lines. He also provided solutions for animal breeding values using

regression (Chapman, 1991). Even though the field of animal breeding and genetics started in the middle of the last century, its development took a long time. The main goal of animal breeding and genetics is the ability to evaluate, select, and rank animals based on their genetic merit. Predictions of genetic merit are usually referred to as breeding values. One of the major breakthroughs in animal breeding is C.R Henderson's mixed model equations (MME) which are used to obtain best linear unbiased predictions of breeding values.

The model is:

$$\mathbf{y = X\beta + Zu + e}$$

y is the vector of observations, **β** is a vector of fixed effects, **u** is a vector of random effects, **e** is a vector of residuals, **X** and **Z** are incidence matrices, and **A** is the relationship matrix.

The MME are written as :

$$\begin{pmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{pmatrix} \begin{pmatrix} \hat{b} \\ \hat{u} \end{pmatrix} = \begin{pmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{pmatrix}$$

Dr Henderson changed the field of animal breeding, his MME methods are still used to this day; he also developed methods to estimate variance components and how to invert the relationship matrix **A** (Henderson, 1976).

The field of animal breeding relies heavily on statistics, there are many complex statistical methods used in animal breeding. Mrode wrote a book entitled "Linear models for the prediction of animal breeding values" that encompasses all the statistical methods from Henderson's mixed model equations MME, BLUP, selection index, Bayesian methods, and estimation of variance components. Computing is a necessary tool in this field, the development

in computer technology has made time consuming analyses simpler and easier. For example, dealing with very large data sets and the inversion of the relationship matrix (A) mentioned earlier which can contain millions of elements when working with large data sets when working with large data sets. We can say that that animal breeding field is a mix of several sciences: biology, statistics, and computer science.

The model used in classical animal breeding and still used to this day is known as the infinitesimal model, which deals with genetic effects as a black box with the assumption that all genes have an effect and therefore, no genes are identified or sequenced. The infinitesimal model has been proven effective and worked excellent.

The model used:

$$\mathbf{y = Xb + Za + e}$$

y is a vector of observations, **b** is a vector if fixed effects, **a** is a vector of random animal effects, **e** is a vector of random residual effects and **X** and **Z** are incidence matrices. MME are used to estimate the effects, and variance components are estimated using methods such as restricted maximum likelihood (REML) and markov chain Monte Carlo (MCMC) methods.

Threshold Models

Traits of interest are not always continuous; some traits are categorical in nature such traits are calving ease, disease status, and many others. Linear and nonlinear models are used, assuming an underlying normal distribution. Gianola (1982) published a paper on theory and implementation of thresholds models. Threshold models are a function of normal distribution for easy use; usually the liabilities follow a normal distribution or a standard normal distribution.

Liabilities are not observable because they are on the underlying continuous scale and on this scale lay the thresholds.

However, the classic animal breeding approach has a few problems and limitations. First, it is not accurate when dealing with secondary traits or lowly heritability traits. Second, for animals with a long generation interval, it can be time consuming.

Quantitative Trait Loci Mapping

Quantitative trait loci (QTL) are genes or stretches of DNA that affect a quantitative trait such as growth, milk production, etc. These traits are controlled by multiple genes scattered throughout the genome. These quantitative traits are very important in the livestock industry because of their impact on profitability. The basis of genomic mapping is the use of linkage disequilibrium, which is a contradiction to Mendel's law of independent assortment. The mapping of these QTLs can be challenging. Several methods have been developed to locate these QTLs. Some of these methods are single-locus model, two locus model, interval mapping, etc.

Zeng (1994) introduced a method for QTL mapping by fitting markers as cofactors. De Koning (2001) presented a method using simple regression to map QTLs in half-sib designs. Regression was used to locate QTLs and estimate their effects and permutation tests were used to obtain empirical threshold values. De Koning (2001) study involved three stages. They started by first identifying candidate genes, then the effects of those genes chosen were estimated using linear regression. Lastly, the phenotypic values were adjusted for the effects of cofactors and linkage groups were analyzed by interval mapping; the results indicated eight QTLs that affect milk yield in Finnish Ayrshire cattle. Yi et al. (2008) proposed a Bayesian approach for QTL mapping known as Bayesian LASSO (Tibshirani, 1996). This Bayesian method simultaneously

fits and estimates the genetic effects of all markers. It also uses prior distributions for the genetic effects that are scale mixtures of normal distributions with mean zero to force some gene effects to have a probability near zero; this method is also referred to as the shrinkage method.

Marker Assisted Selection

Usually selection is based on the available phenotypic information. The development and advances in genotyping and in the field of molecular genetics in general have given rise to a new and precise method of selection, marker-assisted selection (MAS). This selection is based on DNA markers that are associated with QTLs and this marker information can be included in BLUP (Fernando and Grossman 1989). There are three types of markers used, markers based on linkage disequilibrium (LD markers), markers based on linkage equilibrium (LE markers) and mutations in the DNA (Dekkers, 2004). The selection based on genes has provided opportunities for better animal selection by directly selecting on genes especially for traits with low heritability; however, this method did not live up to the expectations (Dekkers, 2004). The development of marker-assisted selection is implemented in three phases: detection, evaluation and implementation phase (Davis et al., 1998). In the detection phase, DNA polymerase is used to detect QTLs and effects of these QTLs are estimated. In the evaluation phase, these markers are used to test QTL segregation in a population of interest. The third phase is implementation where markers are used to test the prediction power of the genetic merit of an animal (Davis et al., 1998).

Microsatellites

Microsatellites are widely used markers in human and animal genetics. Microsatellites, also known as SSRs are sequences of DNA base pairs usually from 1 to 6 base pairs.

Microsatellites are extensively used in human genetic maps due to the multiple polymorphisms they can show. They are scattered across the genome and are a useful tool to associate to traits or diseases. PCR of microsatellites transformed the way of constructing genetic maps (Litt and Lutty 1989; Weber and May, 1989).

Single Nucleotide Polymorphism Discovery

Single nucleotide polymorphism is a variation in DNA sequence, specifically a variation of one nucleotide. For example, in humans when comparing two human genomes, they are 99.9% identical (Cooper et al., 1985). Only certain SNPs have an effect and usually that depends on the location of the SNP in the genome. The majority of SNPs occur in the non-coding region of DNA; only a small percentage of SNPs are causative. SNP genotyping and detection has developed in recent years. SNP detection methods are now robust, less time consuming, and more accurate. SNP detection relies mainly on DNA sequencing or on using denaturing high performance liquid chromatography (dHPLC) (Kwok, 2003).

Genome-wide association study (GWAS)

One important application of SNPs is genome-wide association study (GWAS). It is a method used to associate genetic variations with a particular trait or disease. GWAS is widely used in animal breeding and genetics. Mai et al. (2010) conducted a genome-wide association study for milk production traits in Danish Jersey cattle using a 50k SNP chip. The objective of this study was to map QTLs for milk production traits by a genome-wide association analysis using a mixed model. They used 1,039 bulls; three indices were evaluated for milk production: milk index, fat index and protein index. Animals were genotyped with the SNP50K Chip. The model used was:

$$y_i = u + s_i + bg_{ij} + e_i \quad \text{(Mai et al. 2010)}$$

y_i is the phenotypic value of the i^{th} individual , u is the overall mean, s_i is the sire effect considered to be random, b is the allele substitution effect, and g is the genotype which coded as 0,1 or 2, which corresponds to the number of minor alleles in a specific SNP genotype. Significance of SNPs was tested by *t*-test. They detected 98 QTLs on 27 BTA, 30 QTLs corresponded to milk index, 50 for fat index, and 18 for protein index. Of the 98 QTLs, 33 QTLs were significant using a genome-wide criterion (Mai et al. 2010).

S. Bolormaa et al. (2011) published a paper entitled "Genome-wide association studies for feedlot and growth traits in cattle". The objective of the study was to associate SNPs with feed efficiency traits such as RFI, and also to estimate SNP effects in order to apply them to other populations. SNP genotyping was performed by using the SNP50k chip from Illumina and a 10k chip from Affymetrix. They used 2 types of data: 3 beef cattle data and Holstein dairy cattle data. For statistical analyses, they used a mixed model and fitting 1 SNP at the time as a covariate. They estimated fixed effects, animal effects, and heritability. The SNPs that were significantly associated with RFI in the 50K SNP panel were also tested for association with 10k SNP panel. 75 SNPs were found significant. In the population of Angus 111 SNPs were significantly associated with RFI and only 27 SNPs were significantly associated with RFI in the two populations (S. Bolormaa et al. 2011).

Genomic Selection

Genomic selection is a new method of selection in the animal breeding and genetics field. It is a type of marker-assisted selection; the markers used are SNPs. However, SNP genotyping is still relatively expensive. The ultimate goal of genomic selection is to accurately predict

breeding values of animals usually referred to as genomic breeding values (GBV). Few efficient methods have been developed to compute high accuracy genomic predictions as explained by a paper by VanRaden (2008). In his paper, a simulation study was conducted with data consisting of 2,967 bulls and 50,000 markers. Genomic predictions were computed by both linear and nonlinear systems of equations. The linear method assumes all genes contribute equally to the genetic variance, and the nonlinear method assumes that some genes will have different contributions to the genetic variance. Reliability increased from 32% using a traditional relationship matrix to 63% using genomic relationship matrix, when using the nonlinear method, resulting in reliability of 66% (VanRaden 2008).

Meuwissen (2003) wrote a paper entitled" Genomic selection: The future of marker assisted selection and animal breeding" explaining the development and the implementation of this method. What makes genomic selection a better tool of selection than phenotypic selection is that the QTL mapping step is not needed, especially when using high density SNP panels. Also it allows breeders to preselect animals that have the genes or chromosome segments of better genetic merit (Shaeffer, 2006). Early selection of animals is crucial, especially in species with long generation intervals. When using high density SNP panels, the whole genome is covered and the SNPs are in linkage disequilibrium with QTLs so that SNP are tagging a potential QTLs. There are two methods currently used in genomic selection: single-step method and multi-step method, both methods have advantages and disadvantages. Both methods rely heavily on linkage disequilibrium (LD) between SNPs and QTLs. LD is usually measured using correlation between two loci, which is dependent on allele frequencies.

Multi-step genomic selection

In the multi-step genomic selection, the statistical model generally used is a mixed linear model:

$$\mathbf{y = 1\mu + \Sigma_i x_i g_i + e_i}$$

Where **y** is a column vector of phenotypic values, **1** is simply a column vector of ones and **μ** is the overall mean, $\mathbf{x_i}$ is the genotype or allele substitution effect for a SNP, $\mathbf{g_i}$ is the genotype which takes 0, 1 or 2 corresponding to the number of minor alleles in the genotype, and **e** is the error. To estimate the effects of genes different methods have been proposed. Meuwissen et al. (2001) explained four different methods to implement multi-step genomic selection: least-squares, BLUP- like approach, Bayes A and Bayes B.

Least-squares: it assumes SNP effects as fixed effects is implemented using a regression approach, least-squares is not the best method to use because one will likely face the problem of more explanatory variables than numbers of observations. Meuwissen et al. (2001) proposed a method to deal with this problem which is to perform a regression analysis for every 1 cM segment i.

BLUP: when using this method, one does not face the problem of more effects than observations as in the least-squares method. Also, we assume that variances of all genes are equal, so we only have one variance to estimate. The model used in BLUP is as follows (Meuwissen et al. 2001):

$$\mathbf{y = \mu 1 + \Sigma_i x_i g_i + e}$$

The gene effects are assumed to be random and their estimates are obtained from Henderson's MME.

Meuwissen et al. (2001), also explained the implementation of genomic selection using Bayes A and Bayes B.

Bayes A: this method is similar to BLUP except that the variances of genes or loci are different, so we have to estimate a variance for each gene. The prior distribution of variances is a scaled inverted chi-square $X^{-2}(v, S)$, where S is the scale parameter and v is the number of degrees of freedom. The posterior distribution is also a scaled inverted chi-square: $post(\sigma^2_{gi}|g_i) = X^{-2}$ (v+ n_i, S+$\mathbf{g_i'g_i}$), Gibbs sampling is used to estimate the effects and variances (Meuwissen et al. 2001).

Bayes B: Bayes B assumes that not all genes or loci have an effect; therefore many loci have no genetic variance. The Bayes B method uses a prior distribution as follows:

$$\sigma^2_{gi} = 0 \quad \text{with probability } \pi$$

$$\sigma^2_{gi} \sim X^{-2} \text{ (v, S) with probability } 1-\pi$$

Gibbs sampling could not be applied to estimate variances of loci since no closed form distribution exists for gene variance given the data, so they used Metropolis Hastings algorithm to estimate the effects and variances.

Single-Step Genomic selection

Single-step genomic selection is based on an enhanced relationship matrix, called a genomic relationship matrix that utilizes genomic information as described by Misztal et al. (2009). The model used is as follows:

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

y is a vector of observations, **b** is a vector of fixed effects and **u** is a vector of random animal effects and e is the residual. The relationship matrix can be modified to $\mathbf{H = A + A_\Delta}$ to account for genomic information and $\mathbf{A_\Delta}$ is the deviation from expected relationships. Matrix G replaces the numerator relationship matrix for the genotyped animals (Misztal et al. 2009). Solving MME

is exactly the same as in traditional mixed models. The construction of the G matrix is explained by P.M VanRaden (2008), the relationship matrix G is **ZZ**'/ [2Σpiqi]. Division by [2Σpiqi] makes **G** analogous to **A**.

A Few Issues in Genomic Selection

Inversion of G matrix

The inverse of G matrix can be time consuming, especially when you have large SNP panels. Several methods have been developed to invert the G matrix.

Population stratification and admixture

Genomic selection is usually performed on purebred animals; however, feedlot performance data are usually collected in populations with different or unknown breed composition. Allele frequencies might be different from one breed to another or a minor allele in one breed can be the major allele in another breed. Therefore, using these data to estimate gene effects to predict breeding values can be biased due to population stratification and admixture (Toosi et al., 2010).

Missing genotypes issue

Usually genomic data contains some missing genotypes and not imputing them may have serious consequences on inferences and statistical validity (Yan et al., 2008). Several methods have been developed to impute these missing genotypes such as Fast-Phase (Scheet et al. 2006) that utilizes linkage disequilibrium. Also, it is based on the idea that haplotypes in a certain population cluster together, so hidden markov chain was implemented (Scheet et al., 2006).

Moreover, a neural network approach has been used to impute missing SNP genotypes and it showed decent accuracy (Yan et al., 2008).

Beef Cattle industry: Feedlots and Feeding Techniques

The US fed cattle industry is divided into two major types of feeding systems, one is large commercial feedlots and the other is small farmer feedlots. The first type is designed as a business for large scale production and making a large profit and therefore is better equipped in terms of management, nutrition, anabolic implants, and health care (T. Field "Beef production and Management Decisions fifth edition). All of these factors have a major influence on animal performance and carcass value as discussed in a paper by Mark et al., (2000).

Days on feed (DOF) is a major component in the feedlot system. If one can minimize DOF while achieving the targeted final weight of the animal and end-product quality, profit would increase. Moreover, DOF influences hot carcass weight (HCW), marbling score (MS), and yield grade (YG). Pyatt et al., (2005) studied some factors affecting carcass value and total profit. The study showed that profit is quadratic in response to an increase in DOF. HCW, MS and (YG) have a linear trend with additional DOF. Researchers have always been interested in assessing and predicting performance of animals. McMeniman et al. (2009, 2010) developed a model to predict dry matter intake (DMI) and final shrunk BW (FSBW) based on sex and initial weight. Predicting DOF will enable feedlot managers to make crucial management decisions. However, to predict DOF several factors have to be taken into account such as carcass traits, health status, diet, and hormonal treatment.

Growth promoting implants

Feeding efficiency and growth rate can be increased by using growth-promoting implants. These anabolic implants are usually growth hormones and have many effects other than increasing growth rate. (Foutz et al., 1997) studied the effects of five anabolic implants including a control (no implant) on steer performance, carcass traits, subprimal yields and longissimus muscle properties. The treatments were: C= no implant (control); S = 20 mg estradiol benzoate+200 mg progesterone (Synovex-S); R= combination of 20 mg estradiol+140 mg of trenbolone acetate (TBA), ST= S+140 mg trenbolone acetate (TBA); STT =S+ finaplix-S. They found that steers receiving R, ST and STT were heavier than controls and had higher average daily gains. Feed efficiency also increased for steers receiving R, ST, or STT.

Dietary Additives

Additives such as ractopamine hydrochloride (RAC) and zilpaterol hydrochloride (ZH) are usually fed to livestock during the last 20 to 40 days in order to increase weight gain, improve feed efficiency, along with other important carcass characteristics according to label claims for Zilmax® (Merck Animal Health). This process functions by increasing protein synthesis and decreasing protein breakdown (Moody et al., 2000). In beef cattle, β-adrenergic compounds have the following effects as shown in Moloney et al. (1991), weight gain increased by 10%, feed consumption decreased by 5%, muscle mass increased by 10%, and fat decreased by 30%. In a paper by Scramlin et al. (2010), they evaluated effects of β-adregnergic agonists: ractopamine hydrochloride and zilpaterol hydrochloride on growth performance, caracass traits, and longissimus tenderness of finishing steers. They used a randomized design with 3 treatments: control, RAC and ZH. They found that both RAC and ZH increased final BW, ADG, feed efficiency and HCW compared with controls. ZH significantly decreased adjusted fat thickness

and KPH; ribeye area increased significantly, and increased cut-out yields when compared with controls.

Role of genetics in feedlot animals

For years, beef producers have been relying on phenotypic selection. However with the availability of genomic information, animal selection and management can be improved by assessing the genetic potential of an animal and therefore provide more accurate results. Genomic information can also be used in predicting fed cattle performance traits. Several QTLs and genes have been reported to have an effect on growth and other carcass characteristics. Snelling et al. (2010) conducted a genome-wide association study of growth traits in crossbred beef cattle using a 50 k SNP panel. They found 231 SNPs associated with QTLs affecting growth traits and 12,425 SNPs throughout the genome were also associated with growth traits. Another study (Schenkel et al., 2005) evaluated the association of reported SNPs in the bovine leptin gene with carcass and meat quality traits in a beef crossbred population. The results showed that two leptin exon 2 SNPs were associated with the following traits: fat, lean yield, and grade fat, and also their interaction had a significant effect on LM tenderness (Schenkel et al., 2005).

Summary

The feedlot industry, and the beef industry in general, can benefit from the methods utilized in animal breeding and genetics especially now with the availability of genomic information and the advances in genomic selection and management. Genotyping is becoming a routine technique; including genomic information in the prediction of performance traits such as feed efficiency, growth, and days on feed that can benefit feedlot managers as these traits relate directly to profit.

REFERENCES

Chapman, A. B. 1991. Jay Laurence Lush 1896-1982: a brief biography. J.Anim Sci. 69:2671-2676.

Cooper, D.N. Smith, B.A., Cook,H.J., Nieman,s.,and Shmidtke, J. 1985 An estimate of unique DNA sequence heterozyosity in the human genome. Hum. Genet. 69: 201-205.

Davis, G. P., S. K. DeNise. The impact of genetic Markers on Selection. 1998. J. Anim. Sci. 76:2331-2339

De Koning, D. J, N.F. Schulman, K. Elo, S. Moisio, R. Kinos, J. Vilkki and A. Maki-Tanila. 2001.Mapping of multiple quantitative trait loci by simple regression in half-sib designs. J.Anim Sci 76:616-622

Dekkers, J. C. M. Commercial application of marker-and gene-assisted selection in livestock: strategies and lessons. 2004. J. Anim. Sci E313:E328

Fernando, R. L. and Grossman, M. (1989) Marker-assisted selection using best linear unbiased prediction. Genetics Selection Evolution.467-477

Foutz, C. P, H. G. Dolezal, T. L. Gardener, D. R. Gill, J. L Henseley, and J. B. Morgan. 1997. Anabolic Implants Effects on Steer Performance, Carcass Traits, Subprimal Yields and Longissimus Muscle Properties. J.Anim. Sci. 75:1256-1265.

Georges, M., et al.1993b Microsatellite mapping of a gene affecting horn development in Bos Taurus. Nat. Genet. 4: 206.

Gianola, D. 1982. Theory and analysis of threshold characters. Journal of Animal. Sci Vol. 54, No. 5.

Henderson C. R. 1976. Rapid method for computing the inverse of a relationship matrix.

Kwok,P. Y. Xiangning, C. Detection of single nucleotide polymorphisms. 2003 Mol. Biol.43-60

Litt, M., Luty J. A. 1989 A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am. J. Hum. Genet 44:397

Mark, D. R., T.C. Schroeder, and R. Jones. 2000. Identifying economic risk in cattle feeding. J. Agribus. 18:331-334

Meuwissen, T. Genomic selection: The future of marker assisted selection and animal breeding. Conference 2003.

Meuwissen, T., B. Hayes, and M. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819-1829.

Misztal, I., Legarra, A., Aguilar, I. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information (2009). *J. Dairy. Sci.*, 92:4648–4655.

Moloney, A. P., P. Allen, R. Joseph, and V. Tarrant. 1991. Influence of β-adrenergic agonists and similar compounds on growth. P 455-513 Growth regulation in farm animals. A. M. Pearson and T. R. Dutson, ed. Elsevier Applied Science, New York, NY.

Moody, D. E., D. L. Hancock, and D, B. Anderson. 2000. Phenethanolamine repartitioning agents. CABI publishing, New York, NY.

Pyatt, N. A., et al.2005. Factors affecting carcass value and profitability in early-weaned Simmental steers: II. Days on feed endpoints and sorting strategies. J. Anim. Sci. 83:2926-2937

R.A. Mrode. Linear models for the prediction of animal breeding values, 2nd edition

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

Schenkel, F. S., S. P. Miller, X. Ye, S. S. Moore, J.D. Nkrumah, C. Li, J. Yu, I. B. Mandell, J. W. Wilton and J. L. Williams 2005. Association of single nucleotide polymorphisms in the leptin gene with carcass and meat quality traits of beef cattle. 2005. J. Anim. Sci 83:2009-2020

Scramlin, S. M, W. J. Platter, R. A. Gomez, W. T. Choat, F. K. McKeith and J. Killefer. 2010. Comparative effects of ractopamine hydrochloride and zilpaterol hydrochloride on growth performance, carcass traits and longissimus tenderness of finishing steers. J. Anim. Sci. 88:1823-1829

Snelling, W. M., M. F. Allan, J. W. Keele, L. A. Kuehn, T. McDaneld, T. P. L. Smith, T. S. Sonstegard, R. M. Thallman and G. L. Bennett 2010. Genome-wide association study of growth in crossbred beef cattle. J. Anim. Sci 88: 837-848

Thomas G. Field, Beef Production and management decisions. fifth edition.pages 200-269 Pearson Prentice Hall

Toosi. A , R. L. Fernando and J.C.M Dekkers. Genomic Selection in Admixed and Crossbred populations. J. Anim Sci. 2010. 88:32-46

VanRaden P. M. 2008. Efficient methods to compute genomic predictions.

VanRaden P. M. Genomic Measures of relationship and inbreeding.

VanRaden, P., G. Wiggans, T. Sonstegard, and C. Van Tassell. 2008. Using genomic data to improve dairy cattle genetic evaluations

Weber, J. L., and P. E. May. 1989. Abundant class of DNA polymorphisms which can be typed using the polymerase chain reaction. Am. J. Hum. Genet. 44:388

Zeng, Z.B. 1994. Precision mapping of quantitative trait loci. Genetics 136:1457-1468

CHAPTER 3

PREDICTION OF ZILMAX STATUS OF FEEDLOT STEERS AND HEIFERS USING GROWTH, CARCASS TRAITS, AND GENOMIC INFORMATION[1]

[1] Hay, E., B. Woodward, R. Rekaya. To be submitted to the Journal of Animal Science

ABSTRACT

Accurate prediction of days on feed (DOF) could be beneficial to feedlot managers in decision making as it relates directly to the profitability of the operation. To accurately predict DOF, several management factors and genetic potential of the animal should be considered. However, some of these factors affecting DOF are not always available for all animals in the feedlot. Such factors include health status, diet, or hormonal treatment. In this study, 2,449 feedlot steers and heifers with known Zilmax treatment status were used to develop a statistical model to predict Zilmax status of a given animal. In order to evaluate the ability of predicting Zilmax status, a threshold model for binary response (treated vs control) was implemented within a Bayesian framework via Gibbs sampler. At the liability scale, the model included molecular (SNP genotypes) and growth information (initial weight, hot carcass weight, backfat, marbling scores and days on feed). Different models (based on the explanatory variable included) were implemented and compared using a five-fold cross-validation approach. The validation results indicated that the model that included DOF, IW and HCW as explanatory variables yielded the highest prediction accuracy of Zilmax status. In fact, the prediction accuracy was 0.72 and 0.77 for controls and treated animals, respectively. However, adding genomic information to the model did not lead to an increase in accuracy. In fact, a decrease in accuracies was observed (0.60 for controls and 0.62 for cases) which could be explained in part by the substantial increase in number of parameters in the model.

Keywords: Zilmax, Prediction, Binary response, SNP

INTRODUCTION

The US beef industry has been changing and adapting to meet increasing consumer demand while providing quality product and remaining competitive. These changes have resulted in the development of several techniques and products, such as different feeding methods, growth promoting hormones, and dietary additives. Zilpaterol hydrochloride (ZH) is a dietary additive; it is the active ingredient in the commercial product Zilmax® (Merck Animal Health). Zilmax is a β-adrenergic agonist, it affects growth and carcass traits (Scramlin et al., 2010) by repartitioning feed energy to increase the amount of muscle instead of fat (Avendano et al., 2006). It is usually fed during the last 20 to 40 days followed by a three day withdrawal period. The label claim of Zilmax is that it increases rate of weight gain, improves feed efficiency, and increases carcass leanness. β-adrenergic agonist binds to a β-adrenergic receptor causing a physiological response it increases accretion of skeletal muscle and decreases accretion of fat (Mersman, 1998). This process functions by increasing protein synthesis and decreasing protein breakdown (Moody et al., 2000). In beef cattle, β-adrenergic agonists have the following effects as shown in Moloney et al. (1991), weight gain increased by 10%, feed consumption decreased by 5%, muscle mass increased by 10%, and fat decreased by 30%.

Days on feed (DOF) is an important component in feedlot operations. The ability to accurately predict DOF can be a powerful tool to feedlot managers in decision making and for staying competitive. Furthermore, it will allow feedlot managers to buy cattle that are more feed efficient while maintaining acceptable meat quality characteristics. Similar feedlot cattle performance prediction research has been done. For example, McMeniman et al. (2010) showed that sex and initial BW can be used to predict dry matter intake (DMI) and final shrunk body weight (FSBW), also Galyean et al., (2010) evaluated the accuracy of using initial body weight

and sex to predict important feedlot cattle traits such as average daily gain (ADG), gain to feed ratio (G: F), and HCW. To predict DOF several factors have to be taken into account: carcass traits, health status, diet, and hormonal treatment. Unfortunately, such information is not always available. This study deals with the issue of limited or absent explanatory variables. Specifically, the objective of this study is to predict Zilmax® status of feedlot steers and heifers using growth traits and genomic information. Including genetic information in our model will explain some of the variation not explained by phenotypic data and should help predict DOF and Zilmax status more accurately. Several QTL mapping and genome wide association studies revealed quantitative trait loci (QTLs) and genes that are linked to feedlot cattle performance traits such as growth and carcass traits (Casas et al., 2003: Snelling et al., 2010).

MATERIALS AND METHODS

In this study, a data set of 2,449 market steers and heifers with known Zilmax status, recorded as binary response, were used to evaluate the prediction ability of the proposed procedure. Of the initial data, only 1,753 animals had genomic information, HCW, IW, and other carcass characteristics recorded. From the 1,753 animals, 402 were controls, meaning not treated with Zilmax, and 1,351were treated animals. Given the extreme imbalance between cases and controls and its known effects on the training of the model, we randomly selected 402 case animals out of the 1,351 available resulting in a data set of 804 animals with equal number of cases and controls. The genomic information used was a panel of 200 SNPs; this panel was generated through candidate gene research genes that have been detected as having an effect on cattle performance. A minor allele frequency of 0.05 restriction was applied reducing the number of SNPs to 106.

Statistical model and implementation

The basic latent variable model for the analysis of binary responses in animal breeding context has been presented for almost thirty years (Gianola, 1982). Its basic idea consists of postulating the existence of unobserved continuous random variables that relate to the observed binary or discrete responses.

Let **y** = (y_1, y_2, ..., y_n)' be a nx1 vector of binary responses for animal (i = 1,2,…,n) as it relates to the an underlying random variable satisfying:

$$y_i = \begin{cases} 1 & \text{if } l_i > T \\ 0 & \text{otherwise} \end{cases}$$

where y_i and l_i are the observed binary response and unknown liability for animal i, respectively, and T is a threshold value. Further, it is assumed that

$$l_i \sim N(\mu_i, \sigma_e^2) \qquad [1]$$

The probability of observing a case (treated animal) is:

$$\begin{aligned} p_i &= pr(l_i > T) \\ pi &= pr(\mu_i + e_i > T) \\ p_i &= pr(e > T - \mu_i) \\ p_i &= 1 - pr(e_i < \mu_i - T) \end{aligned} \qquad [2]$$

where Φ is the cumulative distribution function of standard normal. It is clear from [2] that it is not possible to infer separately μ_i, *T,* and σ_e^2. Hence, some restrictions are placed in two of the three model parameters. A common choice is to set T=0, and σ_e^2=1, leading to:

$$p_i = pr(l_i > T | \mu_i) = 1 - \Phi(-\mu_i) = \Phi(\mu_i) \qquad [3]$$

where μ_i can be linearly related to a set of systematic and random effects.

Furthermore, a linear model can be used to express the relation between liability and μ_i. In matrix notation, the model can be written as:

$$\mathbf{l} = \mathbf{X\beta} + \mathbf{e}$$

where **l** is a vector of unobserved liabilities; **β** is the vector of explanatory variables. Depending of the models, the vector **β** included IW, DOF, HCW, BFAT, MS and SNPs; **e** is the vector of residual terms, and **X** is known incidence matrix with the appropriate dimensions.

Based on the assumptions made earlier, the conditional distribution of liabilities given the model parameters was:

$$p(\mathbf{l}|\mathbf{\beta}) \sim N(\mathbf{X\beta}, \mathbf{I}) \qquad [4]$$

For a full Bayesian implementation of [4], prior distributions for the model parameters are required. To avoid potential improper posterior distribution, the following priors were assumed:

A normal distribution with mean zero and a large variance was assumed as prior for the systematic effects **β**.

$$p(\beta_i) \sim N(0, 10^4) \qquad [5]$$

The augmented joint posterior distribution is obtained as the product of density in [4] and [5] and all conditional posterior distributions follow easily. Albert and Chib (1993) and Sorensen et al. (1995) give all needed conditional distributions. The full conditional posterior distributions were all in closed form being normal for element of β, and truncated normal for each l_i, and

scaled inverted chi-square for σ_u^2, σ_s^2 and σ_p^2. A detailed derivation of these conditional distributions can be found in Rekaya et al. (2000).

Convergence diagnostics were assessed using the method of Raftery and Lewis (1992) and the visual inspection of parameters trace plots. The required length of the burn-in period was always less than 6,000 iterations for all parameters. Thus, a total single chain length of 100,000 iterations of the Gibbs sampler was carried out with a conservative burn-in period of 25,000 iterations. The remaining 75,000 iterations were retained without thinning for post-Gibbs analysis.

RESULTS AND DISCUSSION

Summary statistics describing the data are given in Table 3.1. Mean DOF for controls is approximately five days higher than mean DOF of treated animals. This result is expected since Zilmax helps feedlot animals gain weight and muscle amount at a faster rate. Furthermore, Figure 3.1 shows that the majority of control animals have more DOF than treated animals with only a few overlapping animals between the two groups. It is more important to consider the trend observed in this figure rather than the absolute value of the difference between mean DOF of the two groups that could be affected by management decisions (animals that have reached their target final weight may have stayed longer in the feedlot). Moreover, from Table 3.1 the mean HCW for controls is lower than cases, however mean backfat and marbling score are higher for controls than treated due to the reduction of fat caused by Zilmax. This result is in concordance with the conclusions in Mersman (1998).

In order to evaluate the ability to impute the Zilmax status of feedlot animals, several models that included a varying number of explanatory variables were considered. The starting point was a full model that includes all available growth information (IW, HCW, DOF, BFAT, and MS) and genomic information (**M1**). Only SNPs with minor allele frequency greater than 0.05 were used (106 SNPs out the total 200 SNPs). The second analysis (**M2**) is similar to the previous one except that only SNPs with a p-value derived based on single marker analyses smaller than 0.05 were considered. Thus, only 17 SNPs were retained and later used in the model for imputation of Zilmax status. The accuracy of predicting controls and cases was 60% and 62% respectively using M1. Although higher than what could be obtained randomly, these probabilities are low, and a large portion around 40% of Zilmax status was not resolved. For M2, the prediction accuracies increased slightly to 66% and 63% for cases and controls, respectively. The small increase in accuracies compared with M1 could be due to the substantial reduction in the number of unknowns in the model as a result of including only 17 SNPs rather 106 SNPs as was the case in M1. Furthermore, no significant changes in accuracies were observed when the genomic information was completely removed from the model and only growth and carcass quality information was considered (Table 3.2). The model with the best accuracies, obtained after trying several combinations of variables to be considered (Table 3.2), included only initial and hot carcass weights and days on feed as explanatory variables. The accuracies based on five-fold cross validation were 77% and 72% for cases and controls, respectively (Table 3.3). In order to investigate the predictive abilities of these variables, we looked at the distribution of IW, HCW, and DOF between the case and control groups (Figures 3.1, 3.2, and 3.3). The increase in accuracy when having only DOF, IW, and HCW as predictor variables could be due in part to the fact that SNPs do not have a direct effect on Zilmax status; the genomic information (SNPs)

is already explained and included in HCW and DOF data, therefore estimating SNP effects only reduces the performance of the model. Further, when including several explanatory variables in a model, the performance of the model drops which is due to the tradeoff that exists between goodness of fit and parsimony. A simple parsimonious model sometimes performs better than a complex model (Marsh et al, 1996). Moreover, these SNPs were generated from QTLs or genes that were reported to have an effect on various cattle performance traits. Furthermore, the small size of our dataset has negatively impacted the accuracy of predictions because we reduced the number of animals from 1,753 to 804 to balance the number of controls and cases. However, 72% and 77% accuracies to predict controls and cases respectively is reasonable and could be used in practical applications. It is without doubt that knowing Zilmax status will increase the prediction of DOF.

Figures 3.4, and 3.5 show the predictive accuracies using different cutoff points for the probability of declaring an observation as case or control. In both cases accuracies increased, as expected with the increase of the probability cutoff point although the rate of increase tends to level off for very high probabilities. Out of 140 observations with $p(y_i = 1 \mid \mathbf{X}, \hat{\boldsymbol{\beta}}) > 0.8$ (0.8 used as cutoff point), 118 observations or 85% were correctly imputed (their true binary response was one). For the 171 observations with $p(y_i = 1 \mid \mathbf{X}, \hat{\boldsymbol{\beta}}) < 0.2$, 154 observations or 91% were correctly imputed. Thus, for a cutoff point of 0.8 for cases (0.2 for controls), the imputation accuracies were around 90%. This is relevant as these observations (311 observations) represent around 40% of the total number of observations (804 observations) in the data file.

CONCLUSIONS

In this study, we developed a simple yet very useful statistical model that will allow us to impute Zilmax treatment status in order to achieve the end goal of effectively predicting DOF which is a crucial component in the US beef industry and directly related to profit. Moreover, it will give a better understanding of the factors affecting DOF. The results in this study indicate that growth information has reasonable information about Zilmax treatment. However, the current accuracies of 72% and 77% to predict controls and cases respectively preclude the direct use of the predicted Zilmax status for the estimation of DOF. However, a model that accounts for the potential misclassification of the predicted Zilmax status in the calculation of DOF could be a reasonable approach and it is being evaluated.

REFERENCES

Albert, J. H, and S. Chib 1995. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association 88: 669-679.

Avendano-Reyes, L., V. Torres-Rodriguez, F. J. Meraz-Murillo, C. Perez-Linares, F. Figueroa-Saavedra, F. D. Alvarez-Valenzuela, A. Correa-Calderon, and P. H. Robinson. 2006b. Effects of two β-adrenergic agonists on finishing performance, carcass characteristics, and meat quality of feedlot steers. J. Anim. Sci. 84:3259-3265.

Casas, E., S. D. Shackelford, J. W. Keele, M. Koohmaraie, T.P. L. Smith, and R. t. Stone. 2003. Detection of quantitative trait loci for growth and carcass composition in cattle. J. Anim. Sci. 81:2976-2983.

Galyean, M. L., N. Dilorenzo, J.P. McMeniman and P. J. Defoor. 2010. Predictability of feedlot cattle growth performance. J. Anim. Sci. 2010-3328

Herbert W. Marsh. 1996. assessing goodness of fit: is parsimony always desirable? The journal of experimental education 364-390.

McMeniman, J. P., L. O. Tedeschi, P. J. Defoor, and M. L. Galyean. 2010. Development and evaluation of feeding-period average dry matter intake prediction equations from a commercial feedlot database. J. Anim. Sci. 88:3009-3017.

Mersmann, H. J. 1998. Overview of the effects of β-adrenergic receptors agonists on animal growth including mechanism of action. J. Anim. Sci. 76:160-172.

Moloney, A. P., P. Allen, R. Joseph, and V. Tarrant. 1991. Influence of β-adrenergic agonists and similar compounds on growth. P 455-513 Growth regulation in farm animals. A. M. Pearson and T. R. Dutson, ed. Elsevier Applied Science, New York, NY.

Moody, D. E., D. L. Hancock, and D, B. Anderson. 2000. Phenethanolamine repartitioning agents. CABI publishing, New York, NY.

Raftery, A.E. and Lewis, S.M. (1992). One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. *Statistical Science*: **7**, 493-497

Rekaya, R., K. A. Weigel, D. Gianola, B. Heringstad, and G. Klemetsdal. 2000. Methods for attenuating bias of variance component estimates in threshold models when herds are small. J. Dairy Sci. 83: 2337–2346

Scramlin, S. M, W. J. Platter, R. A. Gomez, W. T. Choat, F. K. McKeith and J. Killefer. 2010. Comparative effects of ractopamine hydrochloride and zilpaterol hydrochloride on growth performance, carcass traits and longissimus tenderness of finishing steers. J. Anim. Sci. 88:1823-1829

Snelling, W. M., M. F. Allan, J. W. Keele, L. A. Kuehn, T. McDaneld, T. P. L. Smith, T. S. Sonstegard, R. M. Thallman and G. L. Bennett 2010. Genome-wide association study of growth in crossbred beef cattle. J. Anim. Sci 88: 837-848.

Sorensen, D., S. Andersen, D. Gianola, and I.R. Korsgaard 1995. Bayesian inference in threshold models using Gibbs sampling, Genetics, Selection, Evolution 27, 229-249.

Table 3.1: Summary statistics

Variable	Zilmax	Mean	St dev	Minimum	Maximum
DOF	CTR	174.67	6.48	160	179
	ZTX	169.65	7.06	160	179
HCW	CTR	434.7	38.4	285	544.1
	ZTX	438.7	41.1	264.1	566.8
BFAT	CTR	1.5	0.5	0.3	3.3
	ZTX	1.3	0.4	0.3	3
MS	CTR	424.85	79.21	220	640
	ZTX	411.1	80.26	200	770
IW	CTR	363.1	24.4	318.2	421.8
	ZTX	368.6	28.9	282.3	442.3

Table 3.2: Prediction accuracies of the different models tested

Predictor variables in the model	Accuracy to predict controls %	Accuracy to predict Cases %
IW, HCW	62	56
DOF, IW, HCW	72	77
DOF, IW, HCW, BFAT, MARB	61	62
IW, HCW , SNPs	57	56
DOF, IW, HCW, SNPs	60	62
DOF, IW, HCW, BFAT, MARB, 106 SNPs	60	62
DOF, IW, HCW, BFAT, MARB, 17 SNPs	63	66

Table 3.3: Accuracy of predicting controls and cases based on a five-fold cross validation when including IW, HCW, and DOF only.

	True Zilmax Status	
Predicted Zilmax status	0	1
0	72%	28%
1	23%	77%

Fig 3.1: Days on feed histogram for control and treated animals

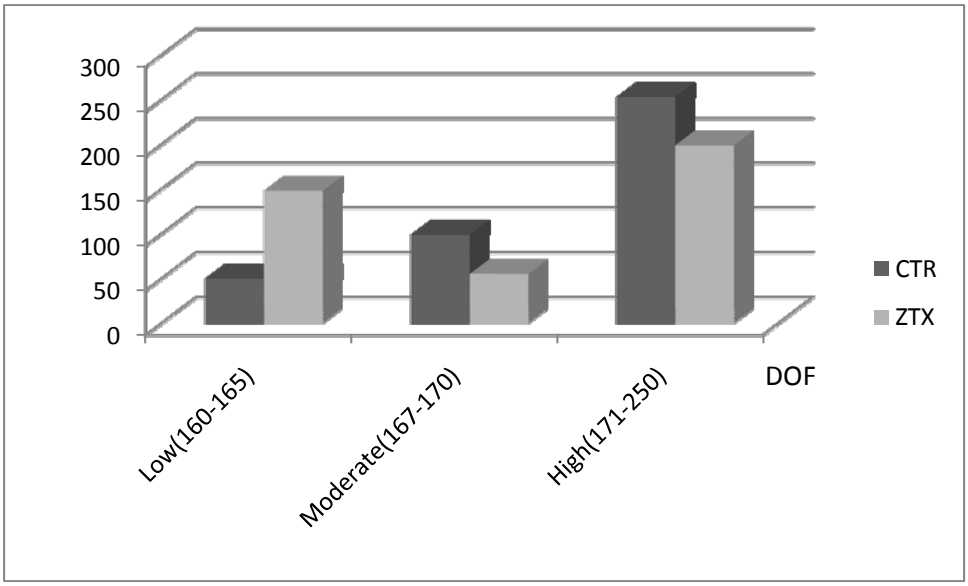


Figure 3.2: Distribution of hot carcass weight (HCW)

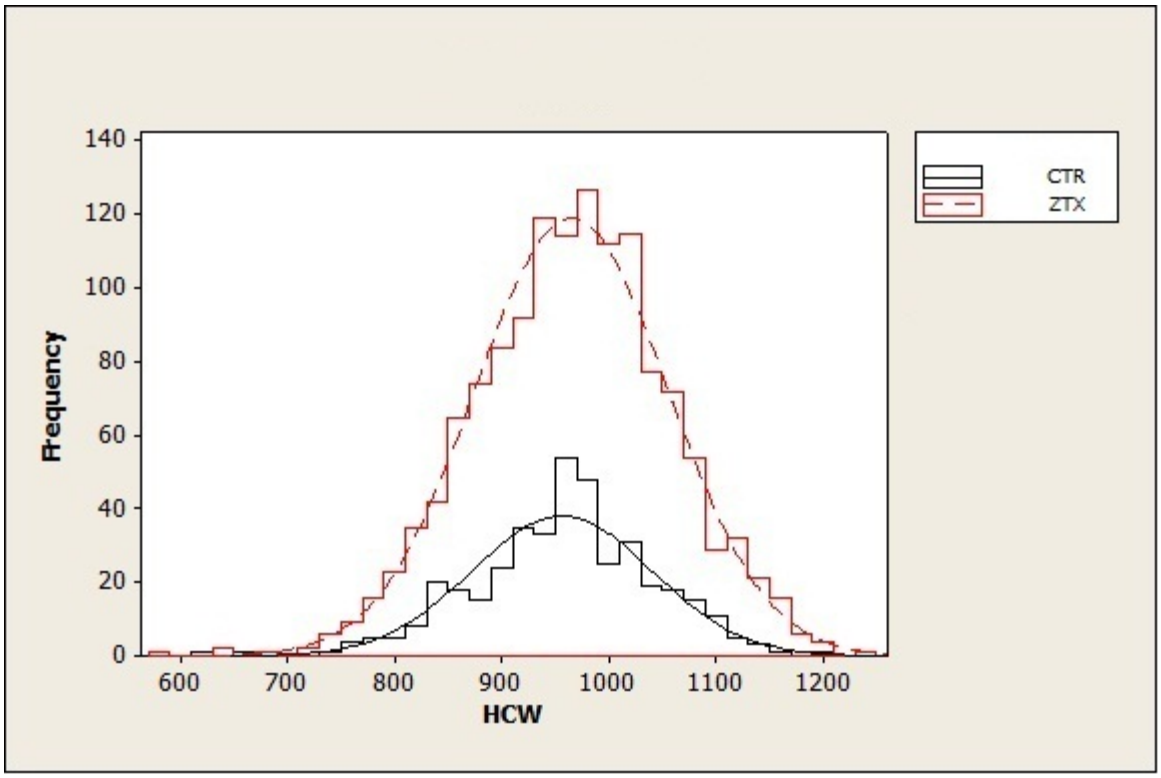


Figure 3.3: Distribution of initial weight

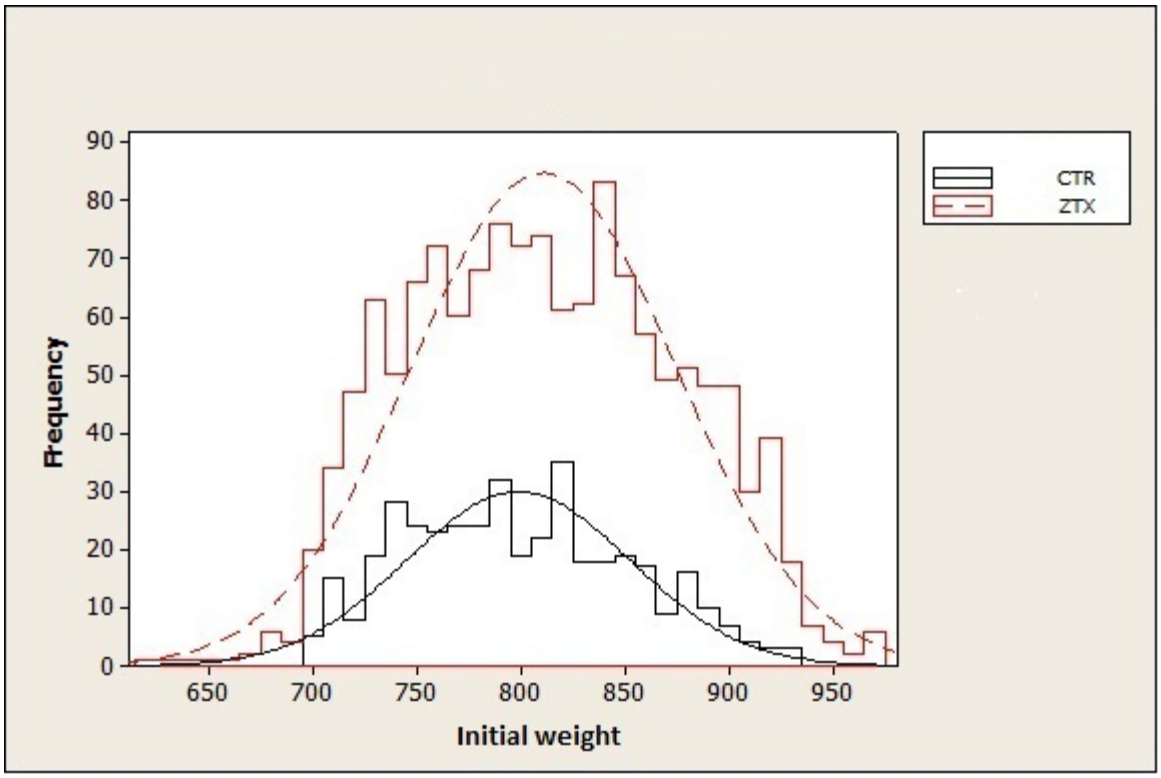


Fig 3.4: Accuracy of predicting controls at different probability classification cutoff points

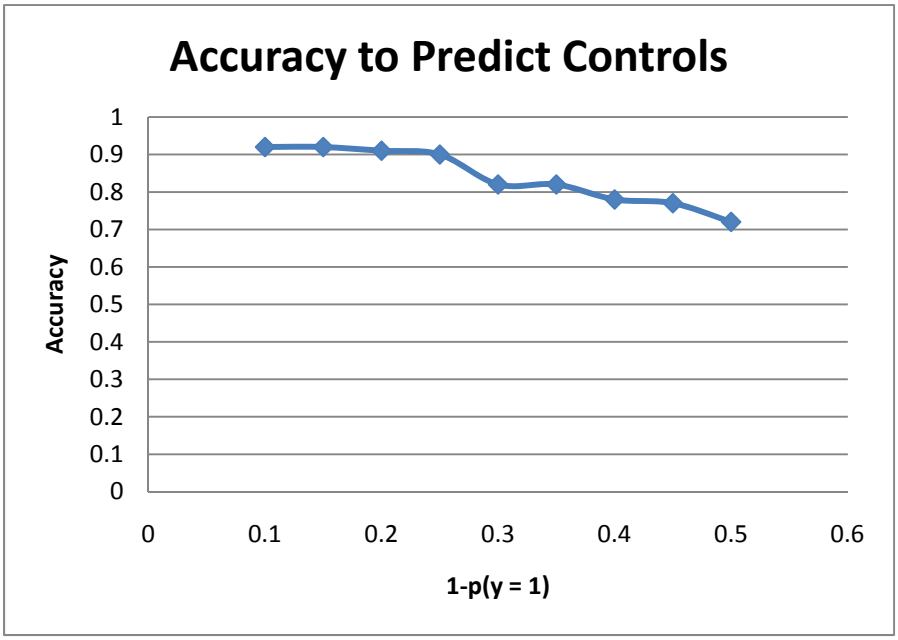
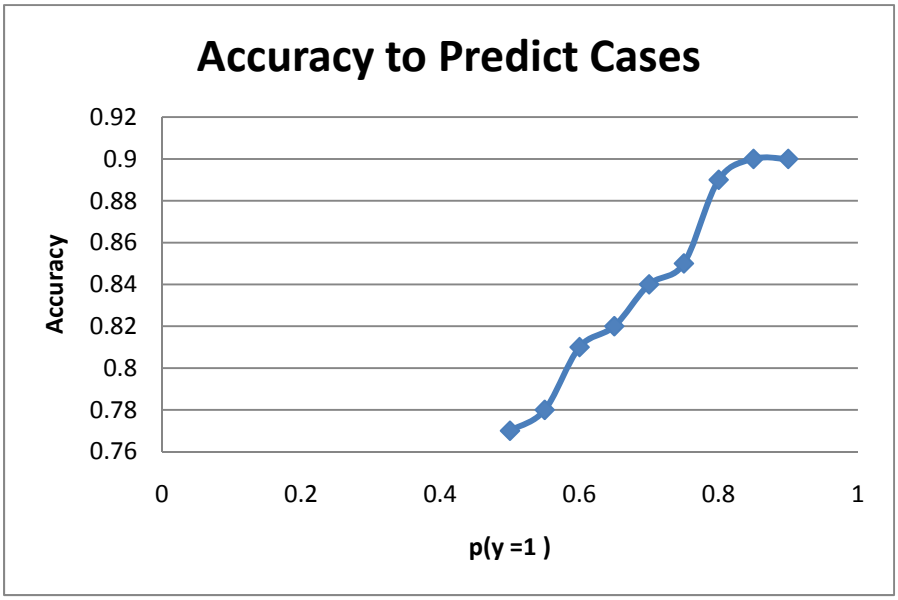


Fig 3.5: Accuracy of predicting cases at different probability classification cutoff points



CHAPTER 4

BAYESIAN ANALYSIS IN THE PRESENCE OF MISSING EXPLANATORY VARIABLES: APPLICATION TO DAYS ON FEED IN FEEDLOT CATTLE[2]

[2] Hay, E, B. Woodward, R. Rekaya. To be submitted to the Journal of Animal Science

ABSTRACT

Days on feed (DOF) is an essential component to profitability in the feedlot industry. Several equations have been developed to predict DOF that include multiple animal and environmental factors. More recently, molecular information (molecular breeding values) for several traits have become available to include in the prediction of DOF. The advantage of this genomic information is its availability as soon as the animal is genotyped for a set of single nucleotides polymorphisms (SNPs). However, important other factors affecting DOF are not always available in all animals in the feedlot. Such factors include health status, drug, or hormonal treatment. In this study, a data set of 10,209 feedlot animals were used to establish a prediction for DOF based on arrival weight, sex, growth, and genotypes for 200 SNP markers, and actual or imputed Zilmax status. Three models were compared for their predictive ability of DOF. The first model (M1) included the following effects: sex, HCW, initial weight, marbling, backfat, ribeye area and 36 SNPs selected out of the available 200 SNPs. The second model, M2, included all effects in the M1 and Zilmax status (case/control) based on the predicted probabilities using 0.5 as a threshold value. The third model, M3, included all effects in M1 and the uncertain Zilmax status. In other words, Zilmax status was not assigned based on the predicted probabilities as in M2 but it was assumed uncertain and to be inferred in the analysis. Additionally, several analyses were conducted where a joint crisp and soft classification of Zilmax status was adopted. The general idea was to assume that Zilmax status is known without uncertainty when the prior probability is below a certain threshold (T1) or greater than a certain threshold (T2). Model comparison criteria (BIC, SSE, and R2) indicated that a statistical model that contemplates uncertainty on Zilmax status are more supported by the data than a model

without Zilmax (M1) or a model that uses crisp classification (M2) based on a fixed probability. Furthermore, it seems that for animals with very low or very high prior probabilities, a crisp classification of Zilmax status of those animals could be helpful as it reduces the number of parameters in the model with little risk of misclassification.

Keywords: Zilmax, prediction, SNPs, DOF, Misclassification

INTRODUCTION

The number of days a feedlot animal is on feed has a major influence on growth and carcass traits of feedlot animals and profit of the feedlot system (Pyatt et al., 2005). Several other factors affect animal performance such as sex, genetics, diet, implants, and health (Mark et al., 2000). Accurately predicting DOF will enable feedlot managers to better assess, and manage feedlot animals. Too long or too short of a feeding period can negatively impact the performance and carcass traits of the animal as well as profit. Pyatt et al. (2005) showed that an increase in DOF resulted in a decrease in profit, moreover average daily gain (ADG) and gain to feed ratio (G:F) decreased with additional DOF. Further, Van Koevering et al. (1995) showed that feed efficiency decreased with additional DOF and that feed efficiency and DOF have a quadratic relationship. Furthermore, DOF is directly dependent on feed efficiency and both of these factors have an impact on profit; 18% increase has been reported in profit as a result of a 10% increase in feed efficiency (Fox et al., 2001). Genetic potential of the animal also has a major effect on feed efficiency, growth, and carcass characteristics. Snelling et al. (2010) conducted a genome-wide association study of growth in crossbred beef cattle, and found several SNPs that affect birth weight (BW) and postnatal growth. Rincon et al. (2009) studied the effect of SNPs in the

STAT6 gene and their association with carcass traits in feedlot animals. Their research showed significant association of a few SNPs with back fat, yield grade, and days on feed. The advantage of using molecular information in the prediction of DOF is its availability as soon as the animal is genotyped for a set of SNPs. The objective of this study was to predict DOF for feedlot steers and heifers using growth traits, management (dietary additive Zilmax), and genomic information. Because some animals in the dataset are missing information about Zilmax status, we included the predicted Zilmax status indirectly in a model that accounts for the potential misclassification of predicted Zilmax status.

MATERIALS AND METHODS

In this study, a dataset of 10,209 market steers and heifers were used to develop a model to predict DOF. Out of the initial 10,209 animals, only 7,719 had genomic information, hot-carcass weight (HCW), initial weight (IW), and other carcass characteristics available. The genomic information consisted of a low density 200 SNP panel; this panel was generated from candidate gene research, meaning genes that were determined to have an effect on cattle performance.

Statistical analyses

In order to evaluate the usefulness of the estimated Zilmax status in the prediction of DOF, several nested models were compared. The first model (M1) included the following effects (sex, HCW, initial weight, marbling, backfat, ribeye area, and 36 SNPs selected out of the 200 SNPs). The second model included all the effects in the M1 and Zilmax status (case / control) based on the predicted probabilities using 0.5 as a threshold value. The third model, M3, included all effects in M1and the uncertain Zilmax status. In other words, Zilmax status was not

assigned based on the predicted probabilities as in M2 but it was assumed uncertain and was inferred in the analysis. The general statistical model for all three analyses could be presented in matrix notation as:

$$p(\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{R}_0) \propto \exp\left[-\frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{R}_0^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right] \qquad [1]$$

where $\mathbf{y}$ is the vector of records, $\mathbf{X}$ is the matrix with the appropriate dimensions relating the observed records to the fixed effects in $\boldsymbol{\beta}$ (with or without Zilmax status), and $\mathbf{R}_0$ is the residual (co)variance matrix.

Additionally, several analyses were conducted where a joint crisp and soft classification of Zilmax status was adopted. The general idea was to assume that Zilmax status is known without uncertainty when the prior probability is below a certain threshold (T1) or greater than a certain threshold (T2). If the probability was smaller than T1 the Zilmax status was assumed zero (control), and if it was greater than T2 the Zilmax status was assumed equal to one (case). For prior probabilities between T1 and T2, a model with uncertain Zilmax status, similar to M3, was assumed. Three values for T1 (0.2, 0.3, and 0.4) and T2 (0.8, 0.7, and 0.6) were tested. Figure 4.1 presents graphically the joint classification of Zilmax status.

For M1 and M2 standard linear regression analyses were implemented. For M3 and the models with joint crisp and soft classification of Zilmax status, a linear regression that imputed Zilmax status based on the phenotypic information was implemented.

Phenotypic information to resolve Zilmax status

For animals with unknown Zilmax status, their phenotypic information or correlated traits could be used to resolve their treatment status. Furthermore, the probability for their predicted Zilmax status as presented in the previous chapter could be used as prior information. This information together with the phenotypic information could resolve the Zilmax status of several animals. Briefly, for each animal with uncertain Zilmax status, two possibilities exist (case or control). Thus, the standardized probabilities for the two possible Zilmaz statuses are given by:

$$p_{i1} = pr(ZTX_i = 1 \mid HCW, ICW, DOF)$$

$$p_{i0} = pr(ZTX_i = 0 \mid HCW, ICW, DOF)$$

where p_{ij} is the probability that the Zilmax status for animal *i* is *j={0,1}* and ZTX_i is the Zilmax status for animal *i*

Let $S_i = \{s_1, s_2\}$ be the set of 2 possible Zilmax statuses for animal i. The only information available in the phenotypic data to discriminate between these two potential configurations is the likelihood of observing the phenotypic record(s) of animal i given each one of the possible Zilmax statuses. Thus,

$$p(\mathbf{y}_i \mid ZTX_i = s_j, \boldsymbol{\beta}, \mathbf{R}_0) \propto \exp\left[-\frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{R}_0^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right] \qquad [2]$$

where ZTX_i is the Zilmax status for animal i, s_j is the j^{th} potential status for animal i, $\mathbf{y}_i$ is the vector of records collected on animal i, $\mathbf{X'}_j$ is the matrix relating the observed records of animal i to the fixed effects in $\boldsymbol{\beta}$, and $\mathbf{R}_0$ is the residual (co)variance matrix. Thus, using the prior information, p_{ij}, the probability of s_j being the true Zilmax status for animal i is given by:

$$p(ZLX_i = s_j \mid \mathbf{y}_i, \boldsymbol{\beta}, \mathbf{R}_0, p_{ij}; j = 1,2) = \frac{p_{ij} \exp\left[-\frac{1}{2}(\mathbf{y}_i - \mathbf{x}_i'\boldsymbol{\beta})'\mathbf{R}_0^{-1}(\mathbf{y}_i - \mathbf{x}_i'\boldsymbol{\beta})\right]}{\sum_{k=1}^{2} p_{ij} \exp\left[-\frac{1}{2}(\mathbf{y}_i - \mathbf{x}_i'\boldsymbol{\beta})'\mathbf{R}_0^{-1}(\mathbf{y}_i - \mathbf{x}_i'\boldsymbol{\beta})\right]} \quad [3]$$

where the denominator of previous equation is the summation of likelihoods, weighted by p_{ij}, for observing the phenotypic record(s) of animal i given each of the possible Zilmax statuses, $s_j\,(j = 1,2)$, for animal i.

In each round of the MCMC algorithm, the true Zilmax status was sampled from a binomial distribution with success probabilities calculated as indicated in equation [3]. At the end of the sampling process, the probability of each possible Zilmax status being the true status for a given animal could be easily computed as:

$$PTS_{ij} = \frac{\text{number of times status } j \text{ was assigned as true for animal i}}{\text{total number of samples}}$$

RESULTS AND DISCUSSION

Data distribution and summary statistics describing the data are shown in Table 4.1 and Table 4.2, respectively. Seventy percent of animals in the dataset have unknown or missing Zilmax status. For animals with missing or unknown Zilmax status, a predicted status has been assigned following the procedure by (Hay et al., 2011). Figures 4.2, 4.3, and 4.4 illustrate the effects of DOF on hot-carcass weight (HCW), backfat, and marbling, respectively. From Figure 4.2, a positive linear trend between HCW and DOF is seen where HCW increased with additional DOF; this result is in concordance with the conclusions in Pyatt et al. (2005). Moreover, marbling increased slightly with additional DOF as shown in Figure 4.4. On the other hand, Figure 4.3 shows a decrease in backfat with additional DOF, this trend seen is due to the reduction of accretion of fat caused by Zilmax as discussed by Mersman (1998).

Table 4.3 presents the comparison between the three models using Bayesian information criterion (BIC), Sum of Square of the Errors (SSE) and the coefficient of determination (R^2). Based on all three criteria, it is clear that M3, which contemplates uncertainty of Zilmax status, is superior to the other two models with substantial reduction in SSE and noticeable increase in R^2. For the other two models, although it seems that M2 (crisp classification of Zilmax status based on prior information) is better than M1 (without Zilmax information) based on SSE and R^2, it is not clear that it is better based on BIC criterion. This conflict in results could be due to the fact that the crisp classification based on p=0.5 has led to substantial errors in the assigned Zilmax statuses, thus eliminating any potential benefit compared to a model without Zilmax.

In order to evaluate the potential benefit of reducing the number of Zilmax statuses to be imputed, especially for animals with very high (pr>T2) or very low prior probability (pr <T1) of

being cases, a set of joint crisp and soft classification criteria were adopted. In these scenarios, if the prior probability of being a case is greater than a threshold value (T2), the Zilmax status was assumed to be a case without uncertainty. However, if the prior probability of being a case is smaller than a threshold value (T1), the Zilmax status was assumed to be a control without uncertainty. For prior probabilities between T1 and T2, the Zilmax status was assumed uncertain and was imputed during the analysis. Using all three model comparison criteria, it seems that the optimum values for T1 is between 0.2 and 0.3 (0.7 and 0.8 for T2). This result seems reasonable because a prior probability of 0.8 indicates high level of certainty about the Zilmax status. Thus, avoiding its imputation will improve the quality of the model as it reduces the number of parameters to be inferred with little risk of making a wrong classification. When T1 was increased to 0.4 and T2 decreased to 0.6, the performance of the model decreased and was worse than a model with complete uncertainty such as M3 (Table 4.4). This results could be in part due to the fact a large proportion of Zilmax statuses assigned based on high T1 (0.4) and low T2 (0.6) are incorrect; thus eliminating any potential benefit from reducing the number of parameters in the model.

The predicted probability for each status, p_{ij}, was used as prior information together with the phenotypic information (DOF) to discriminate between the two possible Zilmax statuses for each animal. For animals with prior probability of being cases greater than 0.8, their associated posterior probability, after inclusion of phenotypic information, increased to 0.93. Similarly, for animals with prior probability of being cases smaller than 0.2, their associated posterior probability decreased to 0.11. These results indicate that phenotypic information was useful in resolving Zilmax status. Only animals with prior Zilmax status greater than 0.8 (group 1) or smaller than 0.2 (group 2) were used because of the high confidence about their statuses. Any

increase in probabilities for group 1 or decrease of probabilities for group 2 will indicate an informative contribution of the phenotypic record.

CONCLUSIONS

The results of this study indicated that a statistical model that contemplates uncertainty on Zilmax status (M3 and M31to M33) are supported more by the data than a model without Zilmax (M1) or a model that uses crisp classification (M2) based on a fixed probability. Furthermore, it seems that for animals with very low or very high prior probabilities, a crisp classification of Zilmax status of those animals could be helpful as it reduces the number of parameters in the model with little risk of wrong classification. Although a grid search was implemented in this study to find the optimum threshold values (T1 and T2), a better and more elegant approach will be to model these thresholds via a change-points model.

REFERENCES

Fox, D.G., P. J. Guiroy, and L. O. Tedeschi. 2001. Determining feed intake and feed efficiency of individual cattle fed in groups. Proc. 33rd Beef Improvement Fed. Mtg., San Antonio, TX.

Mark, D.R., T. C. Shroeder, and R.jones.2000. Identifying economic risk in cattle feeding. J. Agribus.18:331-334.

Mersmann, H. J. 1998. Overview of the effects of β-adrenergic receptors agonists on animal growth including mechanism of action. J. Anim. Sci. 76:160-172.

Pyatt, N. A., L.L. Berger, D. B. Faulkner, P. M. Walker, S.L. Rodriguez-Zas. 2005. Factors affecting carcass value and profitability in early-weaned Simmental steers: II. Days on feed endpoints and sorting strategies. J. Anim. Sci. 83:2926-2937.

Rincon, G., E. A. Farber, C. R. Farber, J. D. Nkrumah and J. F. Medrano. 2009. Polymorphisms in the STAT6 gene and their association with carcass traits in feedlot cattle. J. Anim. Genetics, 878-882.

Snelling, W. M., M. F. Allan, J. W. Keele, L. A. Kuehn, T. McDaneld, T. P. L. Smith, T. S. Sonstegard, R. M. Thallman and G. L. Bennett 2010. Genome-wide association study of growth in crossbred beef cattle. J. Anim. Sci 88: 837-848.

Van Koevering, M. T., D. R. Gill, F.N. Owens,H. G. Dolezal, and C. A. Strasia.1995. Effect of time on feed on performance of feedlot steers, carcass charcacteristics, and tenderness, and composition of longissimus muscles. J. Anim. Sci. 73:21-28.

Table 4.1: Days on feed Data distribution

	Heifers	Steers	Total
Missing ZTX	1643	6076	7719
Known ZTX	0	1753	1753

Table 4.2: Summary statistics

Variable	mean	SD	max	min
DOF	177	39	208	81
IW	304.5	50	507.3	133.6
HCW	351.8	40.5	502.7	206.8
REA	12.8	1.64	21.3	8
MS	412	79	830	190
BFAT	1.2	0.4	3.4	0

Table 4.3: Performance of different models for predicting DOF

Model	BIC	SSE	R^2
Without ZTX (**M1**)	19783.2	507.5	0.56
With predicted ZTX status (1or 0) (**M2**)	19874.7	411.4	0.65
With predicted ZTZ probability (**M3**)	19438.1	379.4	0.69

BIC: Bayesian information criterion; SSE: sum of square of the errors; R^2 is the coefficient of determination

Table 4.4: Performance of models with different thresholds for crisp and soft classification of Zilmax status for predicting DOF

Model	BIC	SSE	R^2
M3	19438.1	379.4	0.69
M31	19407.5	374.7	0.69
M32	19383.2	368.9	0.70
M33	19462.3	385.2	0.69

M3: Using predicted ZTX probabilities (no crisp classification); M31: Crisp classification if probability smaller than 0.2 or greater than 0.8; M32: Crisp classification if probability smaller than 0.3 or greater than 0.7; M33: Crisp classification if probability smaller than 0.4 or greater than 0.6.

Figure 4.1 : Joint crisp and soft classification of Zilmax status

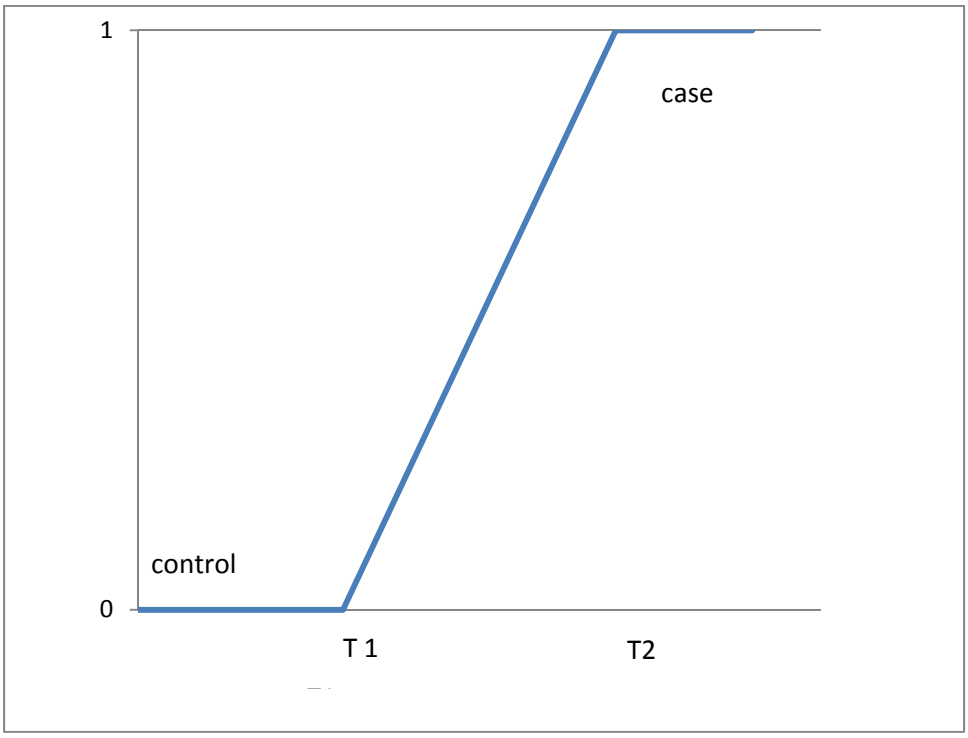


Figure 4.2: Scatter plot of HCW (kg) vs. DOF (days)

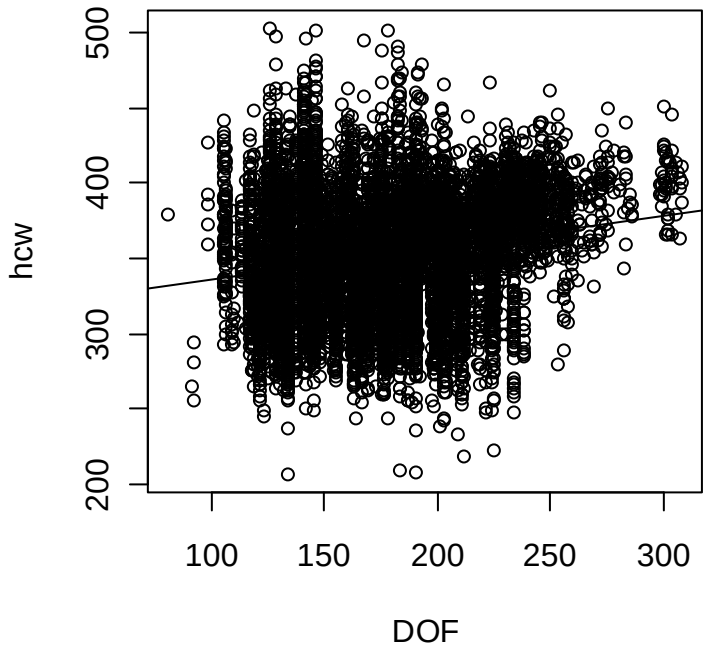


Figure 4.3: Scatter plot of backfat (cm) vs. DOF (days)

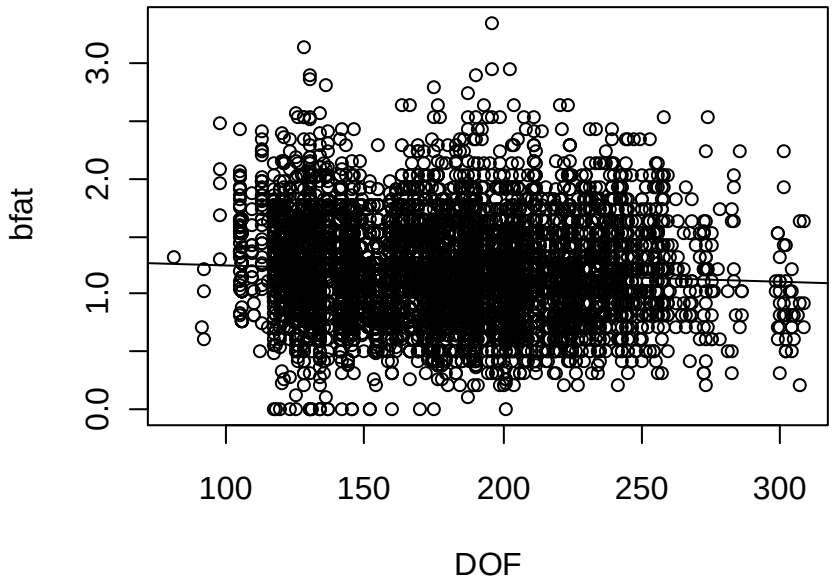
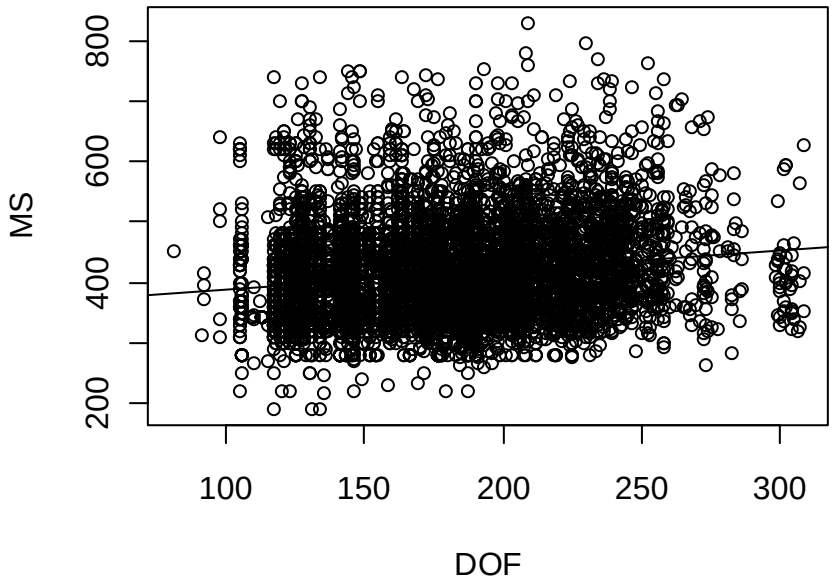


Figure 4.4: Scatter plot of marbling score (MS) vs. DOF (days)



CHAPTER 5

CONCLUSIONS

As discussed earlier, DOF plays a vital role in the feedlot industry. The results of this study showed that DOF can be accurately predicted if accounting for certain factors. We have developed a statistical model that imputes Zilmax treatment status for the purpose to include this important piece of information in the prediction of DOF, and to predict DOF by including the imputed Zilmax status with uncertainty. The results in this study indicated that growth has reasonable information about Zilmax treatment. However, the resulting accuracies 72% and 77% to predict controls and cases, respectively, precluded the direct use of the predicted Zilmax status for the estimation of DOF. Therefore a model that accounts for the potential misclassification of the predicted Zilmax status in the calculation of DOF was implemented and performed better than models not containing Zilmax status. The results indicated that a statistical model that contemplates uncertainty on Zilmax status (M3 and M31to M33) are supported more by the data than a model without Zilmax (M1) or a model that uses crisp classification (M2) based on a fixed probability.