**J. Dairy Sci. 89:3681–3689**

# Random Regression Models for Male and Female Fertility Evaluation Using Longitudinal Binary Data

**T. Averill,* R. Rekaya,†[1] and K. Weigel‡**
*National Agriculture Statistics Service, USDA, Concord, NH 03302-1444
†Department of Animal & Dairy Science, The University of Georgia, Athens 30602-2771
‡Department of Dairy Science, University of Wisconsin–Madison 53706-1284

## ABSTRACT

A longitudinal Bayesian threshold analysis of insemination outcomes was carried out using 2 random regression models with 3 (Model 1) and 5 (Model 2) parameters to model the additive genetic values at the liability scale. All insemination events of first-parity Holstein cows were used. The outcome of an insemination event was treated as a binary response of either a success (1) or a failure (0). Thus, all breeding information for a cow, including all service sires, was included, thereby allowing for a joint evaluation of male and female fertility. An edited data set of 369,353 insemination records from 210,373 first-lactation cows was used. On the liability scale, both models included the systematic effects of herd-year, month of insemination, technician, and regressions on age of service sire and milk yield during the first 100 d of lactation. The random effects in the model were the 3 or 5 random regression coefficients specific to each cow, the permanent effect of the cow, and the service sire effect. Using Model 1, the estimated heritability of an insemination outcome decreased from 0.035 at d 50 to 0.032 at d 140 and then increased continuously with DIM. The genetic correlations for insemination success at different time points ranged from 0.83 to 0.99, and their magnitude decreased with an increase in the interval between inseminations. A similar trend was observed for heritability and genetic correlations using Model 2. However, the average estimate of heritability was much higher (0.058) than those obtained using Model 1 or a repeatability model. In addition, the estimated genetic correlations followed the same trend as Model 1, but were lower and with a higher rate of decrease when the interval between inseminations increased. The posterior mean of service sire variance was 0.01 for both models, and permanent environmental variance was 0.05 and 0.02 for Models 1 and 2, respectively. Model comparison based on the Bayes factor indicated that Model 1 was more plausible, given the data.

**Key words:** longitudinal binary data, random regression, fertility, dairy cow

## INTRODUCTION

In the last 3 decades, the dairy industry in the United States has experienced substantial changes. Royal et al. (2000) reported that in 20 yr (1975–1997) conception rates in Holstein cows had decreased by 0.45% per year (Butler and Smith, 1989; Beam and Butler, 1999). The decrease in fertility can be partially attributed to unfavorable genetic relationships between fertility and production traits (Freeman, 1984; Ranberg et al., 2003). Although several traits are being used for genetic evaluation of reproductive performance, the outcome of an insemination event seems to be the trait of choice, because it is measured early in the breeding season and hence has less environmental influence (Averill et al., 2004). Similar to many other reproductive traits, the outcome of an insemination event depends on both male and female fertility.

Although some attempts have recently been made to jointly analyze male and female fertility (Jamrozik et al., 2005), these 2 traits have been analyzed separately in the majority of cases. Further, not all insemination events during a breeding season are being considered. In fact, traits such as outcome of first insemination, based on the nonreturn rate or conception rate, use only one record per cow within lactation. Thus, additional breeding records of cows having more than one insemination are not considered, leading to information loss. With one record per cow, information about male fertility in second and later services is also lost. Consequently, the resulting service sire effects and female fertility could be biased. To account properly for service sire effect and female fertility in a joint analysis, all insemination events could be considered, such as by using a longitudinal threshold model.

In dairy cattle, random regression models have primarily been used to analyze production traits (Marti and Funk, 1994; Veerkamp and Goddard, 1998; Swalve,

---

Received July 11, 2005.
Accepted January 31, 2006.

[1] Corresponding author: rrekaya@uga.edu

2000; Jensen, 2001) and health-related traits such as mastitis (Kadarmideen et al., 2000; Heringstad et al., 2001; Heringstad et al., 2003; Rekaya et al., 2003). Several random coefficient models, including random regression and splines, have been implemented and compared (White et al., 1999; Pool et al., 2000). The basic idea underlying all these models consists of modeling the additive genetic values (or other random effects in the model) as a function of an observed dependent variable (i.e., time, weight) through a set of random coefficients. Compared with cross-sectional models, the theoretical and biological advantages of using random regression-based models for longitudinal data are numerous and have been reported extensively in the literature (Meyer, 1998; Huisman et al., 2002).

Kirkpatrick and Heckman (1989), among others, have discussed using longitudinal models for the analysis of infinite-dimensional characters. Using dairy cattle data, Rekaya et al. (1998) have applied a longitudinal threshold model for the analysis of sequential binary responses. Further, Veerkamp et al. (2001) have used a longitudinal binary approach to analyze censored survival data via random regression models. In addition to the advantages of using the random regression model for continuous data, longitudinal threshold models offer the possibility of computing quantities of interest to animal breeders that could not be obtained using cross-sectional analyses, such as the probability of observing a success or failure within a specific period. Heringstad et al. (2001) applied the proposed model to analyze the incidence of mastitis in the Norwegian Red dairy cattle population. The estimated heritability ranged from 0.01 to 0.18, with the maximum being by the middle of the lactation. Furthermore, quantities were developed, including the expected number of mastitis episodes per lactation, the expected number of days without mastitis, and the probability of having at least one mastitis episode during a given period of time, and were used as alternatives to a single breeding value for sire selection. Kadarmideen et al. (2001) used a similar approach to analyze mastitis data in the UK Holstein population. Jakobsen et al. (2003) used a bivariate longitudinal threshold-continuous model for analysis of health and production traits in the Danish dairy population. Their model yielded interesting results, such as a description of the susceptibility to diseases as a function of the shape of the lactation curve and the relationships among timing and level of peak milk production and the incidence of mastitis.

In this study, the repeatability threshold model proposed by Averill et al. (2004) for analyzing insemination outcomes in first lactation, where all insemination events of a cow in a breeding season were considered as repeated measurements, is extended via a random coefficient-based model with specific objectives: 1) to use all available breeding information for joint evaluation of male and female fertility; 2) to implement and compare 2 random coefficient functions for modeling the additive breeding value using Bayes factors; and 3) to develop new selection criteria other than the single breeding value.

## MATERIALS AND METHODS

The data consisted of insemination records from first-parity dairy cows generated between 2002 and 2003, provided by AgriTech Analytics (Visalia, CA). The trait of interest was the outcome of all insemination events, where 1 was defined as a successful insemination event and 0 was defined as an unsuccessful (failed) insemination event. Data editing consisted of keeping herd-year contemporary groups and technicians with at least 5 and 10 records, respectively. All extreme-case problem classes for contemporary groups and technicians were removed. Further, cows with inconsistent identification, more than one successful insemination per lactation, or an unrealistic interval between consecutive inseminations were removed. After editing, the data consisted of a total of 210,373 cows with 369,353 records, averaging 1.76 inseminations per cow. The data also included a total of 1,582 technicians, 3,210 service sires, and 967 herd-year contemporary groups.

The pedigree had a high proportion of missing or unknown sires because of the high percentage of cows in California with missing sire identification. After matching the identification numbers in the data and pedigree files provided by AgriTech Analytics with sire pedigrees from the National Association Animal Breeders, international sire identification numbers were determined and used in building the pedigree. The pedigree file consisted of 218,706 animals.

Before describing the longitudinal setting, a basic latent variable model for a cross-sectional binary response is described. Assume the observed binary response, $y_i$, is related to a continuous underlying variable $l_i$ satisfying the following condition:

$$y_i = \begin{cases} 1 & \text{if } l_i > T \\ 0 & \text{if } l_i \le T \end{cases} \qquad [1]$$

where $l_i \sim N(\mu_i, \sigma_e^2)$ and $T$ is a threshold value. The probability of observing a successful insemination (i.e., $y_i = 1$) is

$$P_i = pr(l_i > T \mid \mu_i) = 1 - pr(l_i \le T \mid \mu_i) \qquad [2]$$
$$= 1 - \Phi\left(\frac{\mathrm{T} - \mu_i)}{\sigma_e}\right),$$