University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

Faculty Papers and Publications in Animal Science | Animal Science Department

1-1-2006

# Success at first insemination in Australian Angus cattle: Analysis of uncertain binary responses

Matthew L. Spangler
*University of Nebraska - Lincoln*, mspangler2@unl.edu

R. L. Sapp
*University of Georgia*

R. Rekaya
*University of Georgia*, rrekaya@uga.edu

J. Keith Bertrand
*University of Georgia*, adshead@uga.edu

Follow this and additional works at: http://digitalcommons.unl.edu/animalscifacpub

Spangler, Matthew L.; Sapp, R. L.; Rekaya, R.; and Bertrand, J. Keith, "Success at first insemination in Australian Angus cattle: Analysis of uncertain binary responses" (2006). *Faculty Papers and Publications in Animal Science.* Paper 792.
http://digitalcommons.unl.edu/animalscifacpub/792

This Article is brought to you for free and open access by the Animal Science Department at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Faculty Papers and Publications in Animal Science by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

# Success at first insemination in Australian Angus cattle: Analysis of uncertain binary responses[1]

**M. L. Spangler,[2,3] R. L. Sapp,[3] R. Rekaya, and J. K. Bertrand**

Animal and Dairy Science Department, University of Georgia, Athens 30602-2771

**ABSTRACT:** Field data from Australian Angus herds were used to investigate 2 methods of analyzing uncertain binary responses for success or failure at first insemination. A linear mixed model that included herd, year, and month of mating as fixed effects; unrelated service sire, additive animal, and residual as random effects; and linear and quadratic effects of age at mating as covariates was used to analyze binary data. An average gestation length (GL) derived from artificial insemination data was used to assign an insemination date to females mated to natural service sires. Females that deviated from this average GL led to uncertain binary responses. Two analyses were carried out: 1) a threshold model fitted to uncertain binary data, ignoring uncertainty (M1); and 2) a threshold model fitted to uncertain binary data, accounting for uncertainty via fuzzy logic classification (M2). There was practically no difference between point estimates obtained from M1 and M2 for service sire and herd variance; however, when uncertain binary data were analyzed ignoring uncertainty (M1), additive variance and heritability estimates were greater than with M2. Pearson correlations indicated that no major reranking would be expected for service sire effects and animal breeding values using M1 and M2. Given the results of the current study, a threshold model contemplating uncertainty is suggested for noisy binary data to avoid bias when estimating genetic parameters.

**Key words:** beef cattle, binary data, fertility, fuzzy logic, threshold model

*©2006 American Society of Animal Science. All rights reserved.* J. Anim. Sci. 2006. 84:20–24

## INTRODUCTION

Beef cattle fertility research has increased recently, in part because of increased information availability. To date, there have been several measures of fertility that have been suggested for genetic evaluation purposes. Donoghue et al. (2004) proposed a binary trait, calving to first insemination, which would evaluate the probability that a calving event would arise because of a pregnancy occurring from first insemination. For natural service (**NS**) matings, first insemination was defined as becoming pregnant during the first 21 d of the breeding season, which would correspond to the first heat cycle of the female. With AI data, the exact day of insemination is known; however, to derive the day of insemination for NS data, an average gestation length (**GL**) obtained from AI data is used. Misclassification is inherent with this procedure, although at an unknown rate, because of variation in GL. The current study was motivated by the belief that this misclassification occurs, and to date, there are no studies addressing this issue in field data.

One proposed way to account for this uncertainty is to use fuzzy logic classification. Sapp et al. (2005) evaluated different methods of analyzing such data in a simulation study. They concluded that a threshold model that disregarded misclassification or uncertainty of binary responses could lead to biased inferences. In fact, based on 10 replicates, the genetic variance was severely biased when the data were analyzed ignoring potential misclassifications. Based on these results, they recommended that uncertain or potentially misclassified binary responses be analyzed using a threshold model that contemplates misclassification. Therefore, the objective of the current study was to apply similar methodology to evaluate the usefulness of a threshold model with fuzzy logic classification to account for potential miscoding of the binary trait success or failure of first insemination in a field data set.

## MATERIALS AND METHODS

### *Data*

The data consisted of NS mating and calving records of first-parity females from Australian Angus herds.

[1]Appreciation is expressed to the Angus Society of Australia for the use of their data and to K. A. Donoghue for data editing assistance.

[2]Corresponding author: e-mail: mspanky@uga.edu

[3]The first and second authors contributed equally to this manuscript.

Received August 26, 2004.
Accepted August 29, 2005.

**Table 1.** Summary of the data used in the analysis

| Item | No. |
|---|---|
| Herd | 396 |
| Year | 13 |
| Month of calving | 12 |
| Service sires | 2,116 |
| Daughter records | 33,099 |
| Sires of daughters with records | 4,187 |
| Animals in pedigree file | 79,205 |

Females having their first mating record between 270 and 625 d of age were retained for analysis. Before editing, there were 36,097 records from first-parity females mated by NS between 1987 and 2000 available in the database. Females whose mating resulted in multiple births (n = 288) or those with incomplete records (n = 460) were removed before analysis. Incomplete records included those in which either the sex of calf was unknown or the mating sire was unknown. Gestation length, computed using AI data, was defined as the difference between the insemination date and the subsequent calving date and was averaged by sex of calf as reported by Donoghue et al. (2004). Average GL (SD) was 279.2 d (5.2 d) and 280.3 d (5.2 d) for female and male calves, respectively. Mating records where GL was >2 SD less than the mean derived from AI data were considered to be outliers and were removed during the editing process (n = 225). All herd groups containing only one record, records from service sires with <3 calves, and all records from herds resulting in extreme category problems were removed from the data set. Finally, all service sire groups resulting in extreme category problems were grouped together in one unknown service sire group. After edits, the final data set included 33,099 first-parity NS mating records representing 4,187 sires. The data structure is presented in Table 1.

### *Analysis of First Insemination Success in Beef Cattle*

There are only 2 sources of information available to ascertain conception, both occurring well after the first 21 d of the breeding season: pregnancy checking by either ultrasound or rectal palpation or a calving event. In this study, the latter is used and the underlying assumption is that if conception occurs, a corresponding calving event will occur as well. This may not be true, as it is possible for a female to conceive, but for some reason she may not carry her pregnancy full-term. The only information available in the current data set was days to calving (**DC**), which was computed as the time elapsed between the introduction of the bull and the subsequent calving date (Johnston and Bunter, 1996). Success or failure at first insemination (conception during the first 21 d; **FIS**) was based on the difference between DC and an average GL, where the average GL differed by sex of calf. If the difference between DC and average GL was ≤21 d, then FIS = 1; otherwise, FIS = 0.

Given the variation in GL between cows, it was possible that some cows had uncertain or miscoded FIS. Furthermore, the variation in GL and the difference between DC and average GL could be used to assess the uncertainty or probability of miscoding for every FIS record. One way to account for this uncertainty would be to use fuzzy logic classification, which uses imprecise propositions based on fuzzy set theory to assign partial membership of a set (Chen and Pham, 2001). In the current study, fuzzy logic classification, based on the binary response of FIS and the difference between DC and average GL, was used to calculate the probability of miscoding at time $t_i$ as described by Sapp et al. (2005). Two analyses were carried out: 1) analysis without consideration for potential misclassification, using a threshold model (**M1**); and 2) analysis accounting for potential misclassification, using a threshold model with fuzzy logic classification (**M2**).

### *Statistical Analysis and Computations*

An animal model was used to investigate 2 methods of analyzing uncertain binary responses for FIS. A linear mixed model at the liability scale, which included systematic effects of herd, year, month of mating effects; linear and quadratic covariates for age at mating; and unrelated service sire, animal, and residual as random effects, was used in the analyses. Threshold models are becoming a standard tool for analysis of discrete data in the field of animal breeding and genetics. Extensive literature on its theoretical basis, implementation, and application has been generated in the last 20 yr (Gianola, 1982; Gianola and Foulley, 1983; Sorensen et al., 1995). More recently, Rekaya et al. (2001) proposed a method for analyzing binary data subject to misclassification using a threshold model. In the current study, an extension of such a method, based on fuzzy logic classification as presented by Sapp et al. (2005), was extended to field data.

### *Threshold Model for Analysis of Uncertain Binary Responses*

A detailed description of the methodology can be found in Rekaya et al. (2001) and Sapp et al. (2005). The threshold concept (Falconer, 1981), as applied to data of this type, assumes that FIS is controlled by an underlying normal variable, commonly called a liability, which causes the observed binary response once the liability reaches a threshold level. Here, the basic idea consists of assuming that the observed binary data $\mathbf{m} = (m_1, m_2,, m_n)'$ are a sample of uncertain (misclassified) binary responses of nonobserved real data $\mathbf{y} = (y_1, y_2,, y_n)'$, where each $y_i$ was Bernoulli, with success probability $p_i$ that was expressed as a function of some systematic and random effects. Uncertainty or misclassification occurred if some $y_i$ was switched (e.g., $y_i = 0$ became

$m_i$ = 1; a 0 was coded as 1). Furthermore, for each observation, an indicator variable $\alpha_i$ [$\boldsymbol{\alpha} = (\alpha_1, \alpha_2,..., \alpha_n)'$] was assumed, which takes the value of 1 if $y_i$ was switched, and $\alpha_i = 0$ otherwise. Following notation by Rekaya et al. (2001), each $\alpha_i$ was assumed to be Bernoulli with success probability $\pi_i$ (probability of misclassification or uncertainty) at time $t$ such that $\mathbf{p}(\alpha_i|\pi_t) = \pi_t^{\alpha_i}(1 - \pi_t)^{(1-\alpha_i)}$.

Consequently, the following relationship between $y_i$ and $m_i$, given $\alpha_i$, could be established as $y_i = (1 - \alpha_i)m_i + \alpha_i(1 - m_i)$. Note that for $\alpha_i = 0$ (no misclassification), $y_i$ and $m_i$ are equal as expected. Furthermore, the likelihood function can be written interchangeably as a function of $y_i$ or $m_i$ (Rekaya et al., 2001; Sapp et al., 2005).

A mixed linear model was used for analysis of the underlying liability of FIS. In matrix notation the model could be written as

$$\boldsymbol{\lambda} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z_s s} + \mathbf{Z_u u} + \mathbf{e},$$

where $\boldsymbol{\lambda}$ was a vector of unobserved liabilities, $\boldsymbol{\beta}$ was the vector of fixed effects, $\mathbf{s}$ was the vector of unrelated random service sire effects, $\mathbf{u}$ was the vector of additive effects, and $\mathbf{e}$ was the vector of residual effects. Furthermore, $\mathbf{X}$, $\mathbf{Z_s}$, and $\mathbf{Z_u}$ were the corresponding incidence matrices with the appropriate dimensions.

Using this same notation, let $\mathbf{m}$ be defined as the vector of observed uncertain FIS responses. Assuming that $\mathbf{y}$, the vector of unobserved true FIS responses, and $\mathbf{X}$ are independent and using the relationship between $y_i$ and $m_i$ given earlier, the joint probability of $\boldsymbol{\alpha} = (\alpha_1, \alpha_2,..., \alpha_n)'$ and $\mathbf{m}$, given $\boldsymbol{\theta} = (\boldsymbol{\beta}', \mathbf{s}', \mathbf{u}')'$ and $\boldsymbol{\pi} = (\pi_{t_1}, \pi_{t_2},..., \pi_{tn})'$, was equal to

$$\boldsymbol{p}(\boldsymbol{\alpha}, \mathbf{m} \mid \boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{n} \pi_{t_i}^{\alpha_i}$$
$$(1 - \pi_{t_i})^{(1-\alpha_i)} \; [p_i(\boldsymbol{\theta})]^{(1-\alpha_i)m_i + \alpha_i(1-m_i)}$$
$$[1 - p_i(\boldsymbol{\theta})]^{(1-[(1-\alpha_i)m_i + \alpha_i(1-m_i)])},$$

where $p_i(\boldsymbol{\theta}) = \Phi_i(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_{si}'\mathbf{s} + \mathbf{z}_{ui}'\mathbf{u})$ was the probability FIS for record $i$. The known row vectors were $\mathbf{x}_i'$, $\mathbf{z}_{si}'$, and $\mathbf{z}_{ui}'$, relating the fixed, service sire, and additive effects to the probability of first insemination success, respectively.

Finally, prior distributions for $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ would complete the Bayesian formulation; however, in some situations, $\boldsymbol{\pi}$ is known or could be inferred from external information. In this study, a fuzzy logic approach was used to determine the vector $\boldsymbol{\pi}$.

If the absolute difference between DC and average GL was <16 d or >26 d, there was no uncertainty about the observed FIS response. Otherwise, the following fuzzy logic functions were used to compute the probability of miscoding at time $t_i$ (see Figure 1):

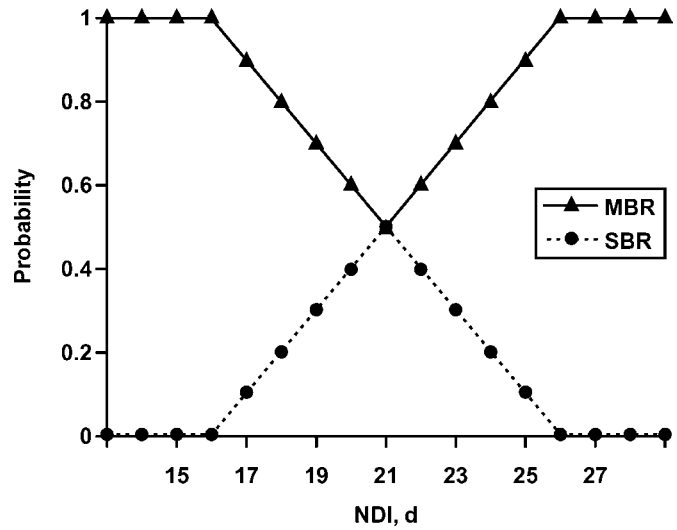


**Figure 1.** Probability that the observed binary response, given the approximate number of days to insemination (NDI), was maintained (MBR) or switched (SBR) using fuzzy logic classification.

$$\pi_{t_i} = 0.5 - (21 - t_i)/10 \; if \; 16 \le t_i \le 21$$

and

$$\pi_{t_i} = 0.5 + (21 - t_i)/10 \; if \; 21 < t_i \le 26.$$

To ensure proper posterior distribution, the following priors were assumed for the parameters in the model:

$$p(\boldsymbol{\beta}_{-h}) \sim N(\mathbf{0}, \mathbf{10^6}) \text{ and}$$
$$p(\boldsymbol{\beta}_h|\sigma_h^2) \sim N(\mathbf{0}, \mathbf{I}\boldsymbol{\sigma}_h^2),$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_{-h}', \boldsymbol{\beta}_h')'$, with $\boldsymbol{\beta}_h$ being the vector of herd effects and $\boldsymbol{\beta}_{-h}$ being the vector of all fixed effects except herd effects, and

$$p(\mathbf{s}|\sigma_s^2) \sim N(\mathbf{0}, \mathbf{I}\sigma_s^2) \text{ and}$$
$$p(\mathbf{u}|\mathbf{A}, \sigma_u^2) \sim N(\mathbf{0}, \mathbf{A}\sigma_u^2),$$

where $\mathbf{I}$ was the identity matrix and $\mathbf{A}$ was a known matrix of relationships between animals. Uniform bounded priors were assumed for $\sigma_s^2$, $\sigma_u^2$, and $\sigma_h^2$.

The joint posterior density is proportional to the product of the density of the conditional distribution times the joint prior density. Samples from the conditional posterior distribution were obtained via Gibbs sampler. After augmentation of the joint posterior with the liabilities (Albert and Chib, 1993; Sorensen et al., 1995), all conditional posterior distributions of model parameters were in closed form and easy to sample from as described by Rekaya et al. (2001) and Sapp et al. (2005). These distributions were normal for the location param-

**Table 2.** Summary of the posterior mean (PM), posterior SD (PS), and lower (HL) and upper (HU) bounds of the high posterior density 95% interval for variance components and heritability

| Parameter[1] | Model[2] | PM | PS | HL | HU |
|---|---|---|---|---|---|
| $\sigma_s^2$ | M1 | 0.135 | 0.009 | 0.117 | 0.154 |
| | M2 | 0.127 | 0.009 | 0.110 | 0.144 |
| $\sigma_h^2$ | M1 | 0.128 | 0.016 | 0.097 | 0.160 |
| | M2 | 0.123 | 0.015 | 0.094 | 0.154 |
| $\sigma_u^2$ | M1 | 0.055 | 0.015 | 0.023 | 0.080 |
| | M2 | 0.031 | 0.009 | 0.015 | 0.048 |
| $h^2$ | M1 | 0.042 | 0.011 | 0.018 | 0.060 |
| | M2 | 0.024 | 0.007 | 0.012 | 0.037 |

[1]$\sigma_s^2$ = service sire variance, $\sigma_h^2$ = herd variance, $\sigma_u^2$ = additive variance, and $h^2$ = heritability, calculated as $\frac{(\sigma_u^2)}{[1 + \sigma_s^2 + \sigma_h^2 + \sigma_u^2]}$.

[2]M1 = analysis of uncertain binary data, without consideration for potential misclassification or uncertainty, using a threshold model; M2 = analysis of uncertain binary data, accounting for potential misclassification or uncertainty, using a threshold model with fuzzy logic classification.

eters, truncated normal for each of the liabilities, binomial for the indicator parameters $\alpha_i$, and scaled-inverted $\chi^2$ distributions for the dispersion parameters. Liabilities were sampled from their truncated normal distribution using an inverse cumulative distribution function technique (Devroye, 1986).

### *Convergence*

Convergence diagnostics were based on the method of Raftery and Lewis (1992) as implemented in the BOA software (Smith, 2003). The required burn-in period was always <3,000 iterations for all parameters in the analyses. Thus, a total chain length of 150,000 iterations of the Gibbs sampler was run with a conservative burn-in of 50,000 iterations. The remaining 100,000 iterations were retained without thinning for post-Gibbs analysis.

## RESULTS AND DISCUSSION

### *Variance Components*

The posterior mean, SD, and the high posterior density 95% [**HPD** (95%)] interval for service sire, herd, and additive variance are presented in Table 2. The results suggest that both service sire and herd variance were less affected by misclassification than additive variance. This finding agrees with the findings of Sapp et al. (2005), who used simulated data to determine that both service sire and herd variances were less affected by misclassification due to the large number of records per service sire or within a given herd. However, point estimates of service sire (0.135) and herd (0.128) variance using M1 were slightly greater than those obtained using fuzzy logic classification (0.127 and 0.123, respectively). The point estimate from the M1 analysis for service sire variance was close to the upper bound of the HPD (95%) interval of the M2 analysis, suggesting that service sire variance was overestimated using M1 compared with M2. The point estimate for additive variance obtained using M1 (0.055) was significantly greater than that from M2 (0.031) because the point estimate (0.055) of M1 fell outside the HPD (95%) interval of M2 (0.015 to 0.048). This result indicates that the additive variance was overestimated when potential misclassification was ignored.

Rekaya et al. (2001) evaluated the possibility of miscoding in dairy cattle for the trait of nonreturn rate at 60 d and found the estimate of additive variance was greater than that obtained from a method accounting for miscoding, although the differences were not as great as in the current study (0.068 and 0.061, respectively). The results of the current study and those of Rekaya et al. (2001) from field data are in contrast to the simulation studies by both Rekaya et al. (2001) and Sapp et al. (2005), which found that when miscoding was ignored in the simulation, the sire variance was underestimated compared with the true value used in the simulation. The conclusion in all these studies was that the estimates of the genetic variance were more accurate when misclassification was accounted for either by modeling the probability of miscoding or by using a fuzzy classification approach. Furthermore, a bias would be expected if misclassification was not considered in the model. The direction of that bias would be unclear and would depend on the specific data and model used; however, the direction of bias may depend in part on the number of 1s and 0s that were wrongly classified.

### *Heritability*

The posterior mean, SD, and HPD (95%) interval for heritability are presented in Table 2. The point estimate of heritability obtained from M1 (0.042) was signifi-

cantly greater than the estimate obtained from M2 (0.024) and fell outside the HPD (95%) interval for M2. This result was expected, given that the additive variance was significantly greater using M1 compared with M2. The HPD (95%) interval using M2 was narrower than the corresponding interval using M1, suggesting more certainty of the estimate obtained from the analysis that accounted for potential misclassification. The results indicated that ignoring potential misclassifications could result in an overestimation of heritability.

### *Pearson Correlations*

Pearson correlations between M1 and M2 for estimated service sire effects and predicted breeding values of the animals in the pedigree file were 0.99 and 0.98, respectively. These results suggest that no major reranking would be expected for either service sire effects or breeding values of animals between the 2 methods. In field data with fewer records per service sire, as may be the case with younger sires, a change in the rank correlation may be anticipated because of limited information with which to infer those effects. Furthermore, it is known that a change in the heritability in a univariate analysis generally does not profoundly affect the ranking of animals. Nonetheless, we expect that the change in the genetic parameters could have an effect on the ranking if the binary trait with potential miscoding is jointly analyzed with correlated traits.

## IMPLICATIONS

The results from this study indicate that differences exist between the 2 methods discussed, particularly for additive variance and heritability, and that in this case, a threshold model that contemplates uncertainty should provide more reliable estimates of these 2 parameters. When estimating parameters for lowly heritable traits, such as those dealing with female fertility where it is intuitive to believe that miscoding occurs, an approach that contemplates the possibility of misclassification may yield more reliable results. Although no reranking of breeding values or service sire effects would be expected given the results of the current study, further research using larger field data sets is warranted to define more completely the benefits of methods that account for classification uncertainty.

## LITERATURE CITED

Albert, J. H., and S. Chib. 1993. Bayesian analysis of binary and polychotomous response data. J. Am. Stat. Assoc. 88:669–679.

Chen, G., and T. T. Pham. 2001. Introduction to Fuzzy Sets, Fuzzy Logic, and Fuzzy Control Systems. CRC Press, LLC, Boca Raton, FL.

Devroye, L. 1986. Non-Uniform Random Variate Generation. Springer-Verlag, New York, NY.

Donoghue, K. A., R. Rekaya, J. K. Bertrand, and I. Misztal. 2004. Genetic evaluation of calving to first insemination using natural and artificial insemination mating data. J. Anim. Sci. 82:362–367.

Falconer, D. S. 1981. Introduction to Quantitative Genetics. 2nd ed. Longman, London, UK.

Gianola, D. 1982. Theory and analysis of threshold characters. J. Anim. Sci. 54:1079–1095.

Gianola, D., and J. L. Foulley. 1983. Sire evaluation for ordered categorical data with a threshold model. Genet. Sel. Evol. 15:201–223.

Johnston, D. J., and K. L. Bunter. 1996. Days to calving in Angus cattle: Genetic and environmental effects, and covariances with other traits. Livest. Prod. Sci. 45:13–22.

Raftery, A. E., and S. Lewis. 1992. How many iterations in the Gibbs sampler? Page 763 in Bayesian Statistics 4. J. M. Bernando, J. O. Berger, A. P. Dawid, and A. F. M. Smith, ed. Oxford Univ. Press, Oxford, UK.

Rekaya, R., K. A. Weigel, and D. Gianola. 2001. Threshold model for misclassified binary responses with application to animal breeding. Biometrics 57:1123–1129.

Sapp, R. L., M. L. Spangler, R. Rekaya, and J. K. Bertrand. 2005. A simulation study for analysis of uncertain binary responses: Application to first insemination success in beef cattle. Genet. Sel. Evol. 37.

Smith, B. J. 2003. Bayesian Output Analysis Program (BOA) Manual, Version 1.0. Univ. Iowa, Iowa City.

Sorensen, D. A., S. Andersen, D. Gianola, and I. Korsgaard. 1995. Bayesian inference in threshold using Gibbs sampling. Genet. Sel. Evol. 27:229–249.