

JOURNAL OF ANIMAL SCIENCE

The Premier Journal and Leading Source of New Knowledge and Perspective in Animal Science

Comparison of methods for handling censored records in beef fertility data: Field data

K. A. Donoghue, R. Rekaya and J. K. Bertrand

J Anim Sci 2004. 82:357-361.

The online version of this article, along with updated information and services, is located on the World Wide Web at:

<http://jas.fass.org/cgi/content/full/82/2/357>



American Society of Animal Science

www.asas.org

Comparison of methods for handling censored records in beef fertility data: Field data¹

K. A. Donoghue,² R. Rekaya³ and J. K. Bertrand

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

ABSTRACT: The purpose of this study was to compare methods for handling censored days to calving records in beef cattle data, and verify results of an earlier simulation study. Data were records from natural service matings of 33,176 first-calf females in Australian Angus herds. Three methods for handling censored records were evaluated. Censored records (records on noncalving females) were assigned penalty values on a within-contemporary group basis under the first method (DCPEN). Under the second method (DCSIM), censored records were drawn from their respective predictive truncated normal distributions, whereas censored records were deleted under the third method (DCMISS). Data were analyzed using a mixed linear model that included the fixed effects of contemporary group and sex of calf, linear and quadratic covariates for age at mating, and random effects of animal and residual error. A Bayesian approach via Gibbs sampling was used to estimate variance components and predict breeding values. Posterior means (PM) (SD) of additive genetic variance for DCPEN, DCSIM, and DCMISS were 22.6d² (4.2d²), 26.1d² (3.6d²), and 13.5d² (2.9d²),

respectively. The PM (SD) of residual variance for DCPEN, DCSIM, and DCMISS were 431.4d² (5.0d²), 371.4d² (4.5d²), and 262.2d² (3.4d²), respectively. The PM (SD) of heritability for DCPEN, DCSIM, and DCMISS were 0.05 (0.01), 0.07 (0.01), and 0.05 (0.01), respectively. Simulating trait records for noncalving females resulted in similar heritability to the penalty method but lower residual variance. Pearson correlations between posterior means of animal effects for sires with more than 20 daughters with records were 0.99 between DCPEN and DCSIM, 0.77 between DCPEN and DCMISS, and 0.81 between DCSIM and DCMISS. Of the 424 sires ranked in the top 10% and bottom 10% of sires in DCPEN, 91% and 89%, respectively, were also ranked in the top 10% and bottom 10% in DCSIM. Little difference was observed between DCPEN and DCSIM for correlations between posterior means of animal effects for sires, indicating that no major reranking of sires would be expected. This finding suggests little difference between these two censored data handling techniques for use in genetic evaluation of days to calving.

Key Words: Beef Cattle, Censored Records, Fertility

©2004 American Society of Animal Science. All rights reserved.

J. Anim. Sci. 2004. 82:357–361

Introduction

Fertility traits have received little attention in genetic evaluations because they are difficult to measure and interpret. Other issues include the best approach for handling animals that fail to calve (censored records). Information for open cows must be included in

evaluations to make best use of data available for reproductive performance. Notter (1988) suggested using threshold theory to assign predicted values to censored females, which was used in several studies (Buddenberg et al., 1990; Meyer et al., 1990). Using threshold theory to generate missing records provided a simple approach to account for the censored nature of days to calving, defined as the time, in days, between when a bull was turned out in the pasture and subsequent parturition (Meyer et al., 1991). More-sophisticated procedures did exist at the time of these studies (e.g., Carriquiry et al., 1987) but were more complex to implement.

Using field data, Johnston and Bunter (1996) generated censored days to calving records by adding a penalty to the largest record within a contemporary group. Sorensen et al. (1998) described a Gibbs sampling scheme using data augmentation for linear mixed mod-

¹Appreciation is extended to the Angus Society of Australia for providing the data; Meat and Livestock Australia for the research scholarship provided to the first author; and D. J. Johnston and C. Teseling for their contributions.

²Present address: Animal Genetics and Breeding Unit, University of New England, Armidale, NSW 2350 Australia.

³Correspondence: Edgar L. Rhodes Center for Animal and Dairy Science (phone: 706-542-0949; fax: 706-583-0274; e-mail: rrekaya@uga.edu).

Received May 2, 2003.

Accepted September 29, 2003.

els for censored traits. In an application of this approach, Donoghue et al. (2004) made random draws from truncated normal distributions to generate censored records on a contemporary group basis. Simulated data were used to compare these two approaches, and model comparison criteria indicated the superiority of the latter approach, although no major reranking of sires was observed.

The objective of this study was to verify results of the simulation study using field data. Three methods for handling censored records were evaluated and compared; assigning penalty value on within-contemporary group basis; simulating records from their respective predictive distributions; and deleting censored records. Correlations between breeding values for sires under all methods were estimated, and rank changes investigated.

Materials and Methods

Data Characteristics

The data consisted of days to calving records from natural service matings of first-calf females in predominantly temperate regions of Australia. The Australian Angus database is a total female inventory recording system, for which mating details on every female in the herd are available. Under this recording system, a disposal code is assigned to any female leaving the herd. Variables recorded in the inventory include animal identification number, date of start of mating, identification number of the mating sire, disposal or fate code, as well as calving date and details of the calf. Further details regarding the inventory can be found in Johnston and Bunter (1996). The trait was defined as the number of days between the time when a bull is turned out in the pasture and the subsequent parturition of the female. Only animals between 270 and 625 d of age when first mated were included in the analysis. Before editing, there were 36,101 trait records from first-calf heifers born between 1987 and 2000 available in the database. Edits performed included removal of 1) animals with incomplete records ($n = 495$), 2) mating records resulting in multiple calves ($n = 288$), 3) single-record contemporary groups ($n = 1,700$), and 4) animals belonging to contemporary groups consisting of only noncalvers ($n = 442$). Incomplete records included records where information on sex of calf and identification of mating sire were missing. After editing, records from 33,176 females were available for analysis. Contemporary group was defined to include animals from the same herd that were mated in the same month and year to the same sire. The sex of calf effect was randomly assigned to either male or female for all animals with censored records. There was 12.5% censoring present in the data, equating to 4,156 (of 33,176) noncalving females. A summary of the data sets can be found in Table 1.

Data Analyses

Penalty Method. The penalty method (**DCPEN**) assigned penalty values to each censored record on a within-contemporary group basis. As suggested by Johnston and Bunter (1996), the highest trait record within each contemporary group was identified, and a constant number (21 d) was added to this record to generate the projected value for all censored records within that group. This constant number is equal to the length of the estrus cycle in cattle, suggesting that females failing to calve would have conceived if given an extra cycle with the bull.

Simulation Method. The simulation method (**DCSIM**) used the approach of Donoghue et al. (2004), and assigned trait values for censored records by simulation from their respective predictive distributions (truncated normal distributions). For all animals in the same contemporary group, the truncation point was the largest observed trait record. The predicted trait value for a censored record was between the truncation point and positive infinity. Thus, an animal with a censored record could not receive a simulated record that was smaller than a noncensored record within her contemporary group. The number of days added to this truncation point for each of the censored records was determined by drawing samples at random from the truncated distribution and depended on the fixed effects in the model, as well as her relationships with other animals.

Missing Method. In order to evaluate the effect of ignoring censored records, a third method (**DCMISS**) was included. Under DCMISS, records for noncalvers were deleted from data set.

Data Analysis. A single-trait mixed linear model was used for analysis of days to calving. In matrix notation, the following model was adopted:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

where \mathbf{y} is a vector of observations, \mathbf{b} is the vector of fixed effects, \mathbf{u} is the vector of additive genetic values of all animals, \mathbf{e} is the vector of residual terms, and \mathbf{X} and \mathbf{Z} are known incidence matrices. The vector \mathbf{b} included contemporary group effects, sex of calf effects, and linear and quadratic covariates for age at mating.

The vector \mathbf{y} includes uncensored data points \mathbf{y}_0 ($m \times 1$) and censored records \mathbf{y}_c ($n - m \times 1$), where n is the total number of observations, such that $\mathbf{y}' = [\mathbf{y}'_0 \mathbf{y}'_c]$.

Augmenting the posterior distribution with the unobserved calving dates corresponding to the censored observations (Tanner, 1996) simplifies the procedure. Let $\mathbf{w} = \{w_j\}$ with $w_j > c_j$, $j = m + 1, m + 2, \dots, n$ and let the augmented data vector be $\mathbf{y}'_A = [\mathbf{y}'_0, \mathbf{w}']$. The joint density of calving dates, both observed and unobserved, given the parameters, is

Table 1. Summary of data characteristics for the penalty (DCPEN), simulation (DCSIM), and missing (DCMISS) methods

Data	Records ^a	Animals ^b	Sires ^c	Herds ^d	CG ^e
DCPEN	33,176	62,857	4,238	470	3,568
DCSIM	33,176	62,857	4,238	470	3,568
DCMISS	29,020	56,797	4,045	459	3,416

^aNumber of animals with trait records.^bNumber of animals in pedigree file.^cNumber of sires of progeny with trait records.^dNumber of herds.^eNumber of contemporary groups.

$$f(\mathbf{y}_A | \mathbf{b}, \mathbf{u}, \sigma_e^2) \propto (\sigma_e^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma_e^2} \left[\sum_{i=1}^m (y_{oi} - \mu_i)^2 + \sum_{i=m+1}^n (w_i - \mu_i)^2 \right] \right\} \quad [1]$$

where y_{oi} is the observed trait record, c_i is the observed value of the trait at censoring time, and $\mu_i = \mathbf{x}_i' \mathbf{b} + \mathbf{z}_i' \mathbf{u}$, ($i = 1, \dots, n$).

Taking into account the augmented data \mathbf{w} , the new vector of unknown parameters of the model will be $\theta'_A = (\mathbf{b}', \mathbf{u}', \mathbf{w}', \sigma_u^2, \sigma_e^2)$. To complete the Bayesian formulation, the following prior distributions were assumed to the unknown parameters:

$$p(\mathbf{b}) \sim N(0, 10^6) \quad [2]$$

$$p(\mathbf{u} | \sigma_u^2) \sim N(0, A\sigma_u^2) \quad [3]$$

where \mathbf{A} is the matrix of additive relationships between animals and σ_u^2 is the additive variance

$$p(\sigma_e^2) \sim U[0, 10,000] \quad [4]$$

$$p(\sigma_u^2) \sim U[0, 1,000] \quad [5]$$

where $U[\cdot]$ is a uniform distribution.

The augmented joint posterior density is easily obtained as the product of distributions in [1] to [5]. Full conditional distributions needed for the implementation of Gibbs sampling can be found in Donoghue et al. (2004). These distributions are normal for the position parameters, truncated normal for the missing records, and scaled inverted chi-squared for the dispersion parameters.

Parameters were drawn from the conditional posterior distributions using Gibbs sampling. The analyses were carried out by the same programs used in Guo et al. (2001) and were developed by R. Rekaya (personal communication). Pearson correlations between posterior means of animal effects for sires were calculated to evaluate the effect of each method on sire ranking. As an additional measure of effect on sire ranking, the percentage of sires in common in the top and bottom

10% of sires ranked with posterior means under different methods was calculated.

Results and Discussion

For all analyses, convergence was assessed using methodology presented by Raftery and Lewis (1992). The required length of the burn-in period was always less than 2,500 iterations for all parameters. Thus, 60,000 iterations of the sampler were run with a conservative 20,000 iterations discarded as burn-in; all remaining 40,000 iterations were retained without thinning for post-Gibbs analysis.

The mean (SD) of days to calving for uncensored females was 302 d (20 d) for DCPEN and DCSIM, and 301 d (19 d) for DCMISS. The corresponding statistics for censored females were 354 d (25 d) and 345 d (23 d) for DCPEN and DCSIM, respectively. Under the DCSIM method, the number of days added to the largest observed trait record within a contemporary group ranged from 5 to 24 d. The majority of censored animals (62%) received a record ranging from 9 to 15 d greater than the largest observed trait record within their contemporary group, whereas relatively few animals (2%) received a censored record equal to the constant number of days (21) assigned under the DCPEN method.

Summaries of the posterior distributions of genetic parameters for days to calving under the different methods of handling censored records are presented in Table 2. Posterior means of the additive variance under DCPEN and DCSIM were similar, and both estimates were within the high posterior density (HPD) interval (95%) of the other method. The corresponding estimate for DCMISS, however, was significantly lower than these estimates and was outside the HPD (95%) interval for both DCPEN and DCSIM. Because the DCMISS method does not consider censored females, thus ignoring an important source of genetic variation, it was expected that this method would produce lower estimates for additive variance. These results show that using either DCPEN or DCSIM to handle censored fertility records had little impact on estimation of additive variance, whereas ignoring censored records (DCMISS) produced significantly lower estimates of this parameter. In the previous simulation study, Donoghue et al.

Table 2. Posterior means, SD, and 95% high posterior density (HPD) intervals for the penalty (DCPEN), simulation (DCSIM), and missing (DCMISS) methods

Method	Parameter	Mean	SD	HPD (95%)
DCPEN	σ_u^2, d^2	22.6	4.2	15.6 to 30.3
	σ_e^2, d^2	431.4	5.0	421.3 to 440.9
	h^2	0.05	0.01	0.03 to 0.07
DCSIM	σ_u^2, d^2	26.1	3.6	19.2 to 33.8
	σ_e^2, d^2	371.4	4.5	362.2 to 380.0
	h^2	0.07	0.01	0.05 to 0.08
DCMISS	σ_u^2, d^2	13.5	2.9	8.6 to 18.7
	σ_e^2, d^2	262.2	3.4	255.6 to 268.8
	h^2	0.05	0.01	0.03 to 0.07

(2003) also found little differences between estimates of additive variance under DCPEN and DCSIM.

Posterior means of the residual variance under all methods were significantly different. In fact, the posterior mean for each method was outside the HPD (95%) interval for the other two methods. The estimate under DCMISS was the smallest of the values, whereas the estimate under DCPEN was the highest value. These results imply that the DCSIM method provides a better fit to the data than DCPEN when censored records are present. These findings follow the same trend evident for this parameter in the simulation study by Donoghue et al. (2004), from which the authors concluded that the higher estimates of residual variance using DCPEN indicated that adding a constant number of days to all censored records within a contemporary group was significantly overestimating the censored records, especially at higher levels of censoring. This suggests that even larger differences between DCPEN and DCSIM in the estimation of residual variance in field data would have occurred had the level of censoring been greater than the 12% observed in the current study.

Posterior means of the heritability under all methods were similar, and all estimates were within the HPD (95%) interval of the other two methods. As a result of the lower estimate of the additive variance under DCMISS, the posterior mean of heritability was slightly smaller than the other methods. The point estimates of heritability in this study for DCPEN and DCSIM are smaller than estimates reported in literature for days to calving or the equivalent trait of calving date for analyses when censored records are included, which ranged from 0.09 to 0.12 (Johnston and Bunter, 1996; Morris et al., 2000; Johnston et al., 2001). However, our estimates were within the range of standard error associated with these literature estimates. The point estimate of heritability for DCMISS was similar and smaller, respectively, to the estimates obtained by Azam and Nielsen (1987) and Meacham and Notter (1987) for calving date with censored records ignored.

Pearson correlations between posterior means of animal effects for sires based on different methods are given in Table 3. As expected, these correlations increased with the number of progeny across all methods.

The correlation for all sires with daughters with records between DCPEN and DCSIM was 0.99, indicating that no major reranking of sires would be expected to occur across these methods. These findings confirm the results reported for ranking of sires in the simulation study by Donoghue et al. (2004). Lower correlations were observed between DCMISS and either DCPEN or DCSIM for all sires with daughters with records (0.77 to 0.81). These results indicate some reranking of sires when censored records are ignored when compared with methods that included noncalving females in the analysis. Assuming that the former approach is inferior, as it is ignoring an important source of genetic variation in fertility, these results highlight the need to include records from noncalving cows in order to accurately estimate differences in fertility for sires. In fact, Notter (1988) noted that inclusion of censored records is required to accurately estimate true sire differences in daughters' fertility, especially for sires that differ markedly in the frequency of censored daughters. A similar trend in correlations was observed for higher accuracy sires (sires with greater than 20 daughters with records) between DCPEN and DCSIM, and DCMISS with DCPEN or DCSIM. The lack of significant differences in the genetic ranking of sires between DCPEN and DCSIM suggests that either censored data handling technique can be successfully used in a genetic evaluation for days to calving.

The percentage of sires in common in the top and bottom 10% ranked on the basis of posterior means of animal effects for sires with daughters with records are given in Table 3. The concordance between evaluations for the top 10% of sires ranged between 60% (DCPEN and DCMISS) and 91% (DCPEN and DCSIM). Concordance between evaluations for the bottom 10% of sires was slightly lower and ranged from 52% (DCPEN and DCMISS) to 89% (DCPEN and DCSIM). These results show that although most sires in the top or bottom 10% ranked similarly under DCPEN and DCSIM, there were exceptions. Further investigation of the sires who ranked outside the top or bottom 10% for the other method, as well as the biggest overall rank changes for all sires, revealed a general trend; a large proportion of their daughters belonged to contemporary groups

Table 3. Pearson correlations and percentage of top and bottom 10% sires in common using posterior means of animal effects for sires with progeny with records for the penalty (DCPEN), simulation (DCSIM), and missing (DCMISS) methods

Comparison	All sires ^a	>20 progeny ^b	Top 10% ^c	Bottom 10% ^d
DCPEN-DCSIM	0.99	0.99	91	89
DCPEN-DCMISS	0.77	0.80	60	52
DCSIM-DCMISS	0.81	0.84	62	56

^aPearson correlations for all sires with progeny with records (n = 4,238 for DCPEN and DCSIM and n = 4,045 for DCMISS with DCPEN or DCSIM).

^bPearson correlations for all sires with more than 20 progeny with records (n = 322 for DCPEN and DCSIM and n = 260 for DCMISS with DCPEN or DCSIM).

^cPercentage of sires in common in the top 10% top sires.

^dPercentage of sires in common in the bottom 10% top sires.

that had a high percentage of censored records. Larger rank changes were observed for sires in the bottom 10% than for the top 10%, suggesting that perhaps the DCSIM method is doing a better job of identifying sires of inferior genetic merit.

Implications

Results for estimation of genetic parameters for days to calving and ranking of breeding values for sires verify the findings of a previous simulation study. Little difference was observed in the heritability of the trait or ranking of sire breeding values between the penalty or simulation method. However, the larger estimate of the residual variance under the penalty method suggests that the simulation approach provides a better method for handling censored records in beef fertility data. The lower estimate of additive variance when censored records are ignored, in conjunction with the expected re-ranking of sires, highlights the unsuitability of this method for evaluation. The lack of significant differences in the genetic ranking of sires between the penalty and simulation methods suggests that either censored data handling technique can be successfully used in a genetic evaluation for days to calving.

Literature Cited

- Azzam, S. M., and M. K. Nielsen. 1987. Genetic parameters for gestation length, birth date and first breeding date in beef cattle. *J. Anim. Sci.* 64:348–356.
- Buddenberg, B. J., C. J. Brown, and A. H. Brown. 1990. Heritability estimates of calving date in Hereford cattle maintained on range under natural mating. *J. Anim. Sci.* 68:70–74.
- Carriquiry, A. L., D. Gianola, and R. L. Fernando. 1987. Mixed-Model analysis of a censored normal distribution with reference to animal breeding. *Biometrics* 43:929–939.
- Donoghue, K. A., R. Rekaya, and J. K. Bertrand. 2004. Comparison of methods for handling censored fertility records in beef cattle data: Simulation study. *J. Anim. Sci.* 82:351–356.
- 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.
- Johnston, D. J., J. M. Henshall, and B. Tier. 2001. Estimate of the genetic correlation between calving success and days to calving in Angus females. *Proc. Assoc. Adv. Anim. Breed. Genet., Queenstown, New Zealand* 14:353–355.
- Meacham, N. S., and D. R. Notter. 1987. Heritability estimates for calving date in Simmental cattle. *J. Anim. Sci.* 64:701–705.
- Meyer, K., K. Hammond, M. J. Mackinnon, and P. F. Parnell. 1991. Estimates of covariances between reproduction and growth in Australian beef cattle. *J. Anim. Sci.* 69:3533–3543.
- Meyer, K., K. Hammond, P. F. Parnell, M. J. Mackinnon, and S. Sivaramasingam. 1990. Estimates of heritability and repeatability for reproductive traits in Australian beef cattle. *Livest. Prod. Sci.* 25:15–30.
- Morris, C. A., J. A. Wilson, G. L. Bennett, N. G. Cullen, S. M. Hickey, and J. C. Hunter. 2000. Genetic parameters for growth, puberty, and beef cow reproductive traits in a puberty selection experiment. *New Zealand J. Agric. Res.* 43:83–91.
- Notter, D. R. 1988. Evaluating and reporting reproductive traits. Pages 21–42 in *Proc. Beef Improv. Fed. 20th Res. Symp. Annu. Mtg., Albuquerque, NM.*
- Raftery, A. E., and S. Lewis. 1994. How many iterations in the Gibbs sampler? Pages 763–773 in *Bayesian Statistics 5*. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, ed. Oxford University Press, New York.
- Sorenson, D. A., D. Gianola, and I. R. Korsgaard. 1998. Bayesian mixed-effects model analysis of a censored normal distribution with animal breeding applications. *Acta Agric. Scand. Section A.* 48:222–229.
- Tanner, M. A. 1996. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. 3rd ed. Springer-Verlag, New York.

References

This article cites 9 articles, 5 of which you can access for free at:
<http://jas.fass.org/cgi/content/full/82/2/357#BIBL>

Citations

This article has been cited by 5 HighWire-hosted articles:
<http://jas.fass.org/cgi/content/full/82/2/357#otherarticles>