A Zero-Inflated Poisson Model for Genetic Analysis of the Number of Mastitis Cases in Norwegian Red Cows

M. Rodrigues-Motta,* D. Gianola,*†‡ B. Heringstad,‡ G. J. M. Rosa,† and Y. M. Chang†

*Department of Animal Sciences and

†Department of Dairy Science, University of Wisconsin, Madison 53706

‡Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, PO Box 5003, N-1432 Ås, Norway

ABSTRACT

The objective was to extend a zero-inflated Poisson (ZIP) model to account for correlated genetic effects, and to use this model to analyze the number of clinical mastitis cases in Norwegian Red cows. The ZIP model is suitable for analysis of count data containing an excess of zeros relative to what is expected from Poisson sampling. A ZIP model was developed and compared with a corresponding Poisson model. The Poisson parameter followed a hierarchical structure, and a residual term accounting for overdispersion was included. In both models, the Poisson parameter was regressed 1) on the year, month, and age at first calving; 2) on the logarithm of the number of days elapsed from calving to the end of first lactation; and c) on herd and sire effects. Herd and sire effects were assigned normal prior distributions in a Bayesian analysis, corresponding to a random effects treatment in a frequentist analysis. An analysis of residuals favored the Poisson model when there were 2 or more cases of mastitis during first lactation, with very small differences between the ZIP and Poisson models at 0 and 1 cases. However, the residual assessment was not satisfactory for either of the models. The ZIP model, on the other hand, had a better predictive ability than the corresponding Poisson model. Posterior means of the sire, herd, and residual variances in the ZIP model (log scale) were 0.09, 0.37, and 0.36, respectively, highlighting the importance of herds as a source of variation in clinical mastitis. The correlation between sire rankings from the ZIP and Poisson models was 0.98. A weaker correlation would be expected in a population with more severe inflation at zero than the present one. The estimate of the perfect state probability p was 0.32, indicating that 32% of the animals would be in the perfect state, either because they are resistant or because they were not exposed to mastitis.

Key words: Bayesian method, mastitis, zero-inflated data

INTRODUCTION

In quantitative genetic analysis, mastitis is often scored as “absent” or “present” and analyzed with linear (e.g., Carlén et al., 2005) or threshold models (e.g., Heringstad et al., 2003a; Chang et al., 2004), with the latter taking the binary nature of the data into account. Cows may have more than one case of mastitis during lactation, and longitudinal binary response models have been used as well (Heringstad et al., 2003b). An alternative form of analysis is one based on counting the numbers of episodes of the disease within a given time period. A candidate sampling model for studying variation in the number of clinical mastitis cases during a lactation is provided by the Poisson distribution (Schukken et al., 1990; Barnouin et al., 2005). Unfortunately, this theoretical distribution is seldom suitable because it introduces a constraint on the variance to mean ratio. The presence of overdispersion leads to incorrect inferences and to inappropriate statistical tests (Gasqui et al., 2000). Variation attributable to Poisson sampling often accounts for only a proportion of the total variation observed (Breslow, 1984; Brillinger, 1986). One way of dealing with overdispersion is to fit a mixed-effects model, such as the pure Poisson mixed model proposed by Foulley et al. (1987), or extend this model to account for “extra-Poisson” variation by introducing a residual term in the structure (Tempelman and Gianola, 1996).

In some situations, overdispersion may be due to an excessive number of zero counts (Lambert, 1992). In such a case, a zero-inflated Poisson (ZIP) model would be suitable. In a ZIP model, it is assumed that, with probability p , the only possible observation is 0 and, with probability $1 - p$, that a Poisson random variable with parameter λ is observed. Here, p is the probability of the “perfect” state (i.e., an animal is either resistant or not exposed to the disease), and λ is the mean number

Received December 27, 2006.

Accepted July 12, 2007.

¹Corresponding author: motta@calshp.cals.wisc.edu

of mastitis cases in the “imperfect” case (i.e., an animal is liable to the disease).

The main objective of this study was to extend the ZIP model to accommodate correlated genetic effects, and to use this model for genetic analysis of the number of mastitis cases. As an illustration, an application was made to data containing information on the number of cases of clinical mastitis in first-lactation Norwegian Red cows from a previously analyzed Norwegian data set from the early 1990s. Further, the ZIP model was compared with a corresponding Poisson model via a residual and predictive ability analysis.

MATERIALS AND METHODS

Data

The data used to implement and test the model consisted of clinical mastitis information on 36,178 first-lactation Norwegian Red cows. Cows in this pilot data set were daughters of 245 sires that had their first progeny test in 1991 and 1992. This data set (same cows) has been used in several other studies (e.g., Chang et al., 2004, 2006). Although it was not the purpose of this study, an advantage of using results from the same cows is that they can easily be compared with results from other studies that used different methods.

Only records of daughters with first calving in 1990 through 1992, and from herds having at least 5 daughters of any of the test bulls, were used. For each cow, the number of cases of veterinary-treated clinical mastitis (NCM) in first lactation, from 30 d before calving to either 300 d after calving or culling, whichever occurred first, was counted. Cows having a second calving before 300 d after first calving were excluded from the analysis. The interval between treatments had to be at least 5 d, to avoid counting repeated treatments of the same case as distinct mastitis episodes. About 76% of the cows did not have a case of clinical mastitis during first lactation, whereas 15.8, 5.1, and 1.6% had 1, 2, and 3 cases, respectively. Only 315 cows had more than 3 episodes of clinical mastitis during first lactation.

Discriminating between simple Poisson and ZIP models can be done by testing the null hypothesis, $H_0: p = 0$. Asymptotic tests proposed by Rao and Chakravarti (1956) and van den Broek (1995) were used as part of an explanatory analysis. The Rao and Chakravarti criterion was calculated as

$$U = \frac{[n_0 - E(n_0)]}{\sqrt{\text{var}(n_0)}} = \frac{n_0 - n \left(\frac{n-1}{n} \right)^{n\bar{y}}}{\sqrt{n \left(\frac{n-1}{n} \right)^{n\bar{y}} - n^2 \left(\frac{n-1}{n} \right)^{2n\bar{y}} + n(n-1) \left(\frac{n-2}{n} \right)^{n\bar{y}}}},$$

where $n_0 = 27,745$ is the number of zeros in the data; $n = 36,178$ is the total number of observations; and $\bar{y} = 0.34$ is the mean number of mastitis cases in the data set. The U statistic is asymptotically distributed as an $N(0, 1)$ variable; the realized value of $U = 62.59$ leads to rejection of H_0 . The score test is given by

$$S = \frac{(n_0 - n\hat{p})^2}{n\hat{p}(1 - \hat{p}) - n\hat{p}^2\bar{y}} = 3917.3,$$

where $\hat{p} = e^{-\bar{y}} = e^{-0.34}$ is the estimate of the probability of having no mastitis cases. The realized value of $S = 3917.3$ was referred to a χ^2 distribution with one degree of freedom, leading to rejection of the hypothesis that p is 0. The tests suggest overdispersion in the number of mastitis cases relative to the homogeneous Poisson sampling, and part of this may be caused by the observed excess of zeros. The tests are suggestive, but not conclusive, because they are based on the assumption of a homogeneous population.

An excess of zeros by level of year, month, and age at first calving was also investigated. For each level of these 3 potential explanatory variables, the observed proportion of zeros was compared against a Poisson distribution with parameter equal to the mean number of mastitis cases in the respective level, as shown in Table 1. An excess of zeros was found for all levels. However, there were some herds and sires without extra zeros according to this criterion. For the sire with the largest frequency of daughters with mastitis, the observed and expected proportions of zeros were 0.64 and 0.84, respectively. A sire pedigree file, with a total of 437 males, was built by tracing the pedigree of the 245 sires with daughters in the data set, through sires and maternal grandsires, as far back as possible.

Statistical Model

A ZIP model was developed for analysis of NCM. Let $\mathbf{y} = (y_1, \dots, y_n)'$ be a vector of NCM records on cows, such that $y_i = 0$ with probability p , and $y_i \sim \text{Poisson}(\lambda_i)$ with probability $1 - p$, with $n = 36,178$. Note that if p is equal to zero, the setting yields a Poisson distribution with mean λ_i . The probability of a cow being healthy is

$$P(y_i = 0) = p + (1 - p)e^{-\lambda_i},$$

and the probability of a cow having k cases of mastitis is

$$P(y_i = k) = \frac{(1 - p)e^{-\lambda_i}\lambda_i^k}{k!}, \quad k = 1, 2, \dots, \infty.$$

Note that p was assumed to be homogeneous, meaning that it is the same for all animals. Let $\lambda =$

Table 1. Observed and expected (under a homogeneous¹ Poisson distribution) proportion of zeros for each level of calving year, calving month, and age at calving

Item		Period						
	Calving year							
Level	1990	1991	1992					
Observed	0.78	0.77	0.75					
Expected	0.77	0.70	0.70					
	Calving month							
Level	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
Observed	0.78	0.76	0.76	0.77	0.74	0.76	0.76	0.77
Expected	0.72	0.70	0.70	0.72	0.65	0.70	0.70	0.72
Level	Sep	Oct	Nov	Dec				
Observed	0.78	0.78	0.76	0.76				
Expected	0.72	0.71	0.70	0.70				
	Age at calving							
Level, mo	<20	20	21	22	23	24	25	26
Observed	0.81	0.78	0.79	0.76	0.77	0.77	0.76	0.76
Expected	0.77	0.73	0.73	0.70	0.71	0.70	0.70	0.70
Level, mo	27	28	29	30	31	32	>32	
Observed	0.76	0.76	0.77	0.75	0.77	0.75	0.79	
Expected	0.70	0.71	0.70	0.68	0.70	0.70	0.73	

¹The Poisson parameter is the mean number of mastitis cases for each level of the respective variable.

$(\lambda_1, \dots, \lambda_n)'$. The conditional (given the fixed and random effects) log-likelihood function for the homogeneous p ZIP model is given by

$$\begin{aligned}
 \log[l(\lambda, p \mid \beta, \mathbf{h}, \mathbf{s}, \mathbf{y})] &= \\
 \log \left\{ \prod_{y_i=0} [p + (1-p)e^{-\lambda_i}] \prod_{y_i>0} [(1-p)e^{-\lambda_i} \lambda_i^{y_i}/y_i!] \right\} \\
 &= \sum_{y_i=0} \log [p + (1-p)e^{-\lambda_i}] + \sum_{y_i>0} \log (1-p) \\
 &\quad - \sum_{y_i>0} \lambda_i + \sum_{y_i>0} y_i \log \lambda_i - \sum_{y_i>0} \log (y_i!),
 \end{aligned}$$

where $l(\lambda, p \mid \beta, \mathbf{h}, \mathbf{s}, \mathbf{y})$ is the conditional likelihood function for the homogeneous p ZIP model. Further, λ and p satisfied the models:

$$\lambda^* = \log(\lambda) = \mathbf{X}\beta + \mathbf{W}\mathbf{h} + \mathbf{Z}\mathbf{s} + \varepsilon; p^* = \log\left(\frac{p}{1-p}\right),$$

where β is a vector of effects of year of first calving (3 levels), month of first calving (12 levels), age at first calving (15 levels; Table 1), and regression on the log of the number of days elapsed from calving to the end of first lactation (culling or d 300); $\mathbf{h}$ is a vector of herd

effects of order 5,286; $\mathbf{s}$ is a vector of sire transmitting abilities of order 437; and $\mathbf{X}$, $\mathbf{W}$, and $\mathbf{Z}$ are the corresponding incidence matrices. Further, ε is a vector of residuals, which was assumed to follow the multivariate normal distribution $N(\mathbf{0}, \mathbf{I}_n\sigma_\varepsilon^2)$, where σ_ε^2 is the residual variance representing all unaccounted for sources of variation. Part of this unaccounted for variation is because a sire model is used here, so three-fourths of the additive genetic variance was not modeled explicitly (Damgaard et al., 2006). The distributions of herd and sire effects were assumed to be independent multivariate normal with mean zero and covariance matrices $\mathbf{I}_{5,286}\sigma_h^2$ and $\mathbf{A}\sigma_s^2$, respectively, where $\mathbf{A}$ is a known matrix of additive genetic relationships between sires, and σ_h^2 and σ_s^2 are variances of herd and sire effects, respectively, on a logarithm scale. The ZIP model was compared with a Poisson model with the same structure for the Poisson parameter. Inferences about unknown parameters were based on the Bayesian approach, via a Markov chain Monte Carlo (MCMC) sampling algorithm.

Prior Distributions. Models were developed hierarchically, with all parameters regarded as random variables following some prior distribution. The prior distribution adopted for λ^* was a normal distribution with appropriate mean and variance σ_ε^2 . Independent

bounded uniform priors were assigned to p^* and to each of the elements of β , that is, $p^* \sim U(-5, 5]$ and $\beta \sim U(-50, 50]$. As stated above, herd and sire effects were assigned independent normal prior distributions with mean zero and variances σ_h^2 and σ_s^2 , respectively. Scale-inverse χ^2 prior distributions were used for the variances of herd, sire, and residual effects, that is, $\sigma_h^2 \sim v_h \delta_h \chi_{v_h}^{-2}$, $\sigma_s^2 \sim v_s \delta_s \chi_{v_s}^{-2}$ and $\sigma_\varepsilon^2 \sim v_\varepsilon \delta_\varepsilon \chi_{v_\varepsilon}^{-2}$, respectively, where the degrees of freedom parameters were set to $v_h = v_s = v_\varepsilon = 5$ and the scale parameters were assigned the values $\delta_h = \delta_s = \delta_\varepsilon = 1$. These values were chosen to achieve vague proper priors for the variance components.

Fully Conditional Posterior Distributions and Sampling Procedure. The fully conditional distributions of individual parameters were deduced by taking all other parameters as fixed, and absorbing them into the integration constant of the conditional posterior distribution of interest. The fully conditional densities of σ_ε^2 , σ_h^2 , and σ_s^2 are scale-inverse χ^2 with appropriate parameters, and the joint fully conditional density of β , $\mathbf{h}$, and $\mathbf{s}$ is a multivariate normal distribution with appropriate mean and (co)variance matrix (Wang et al., 1993, 1994).

The fully conditional density of each element of λ^* and the density of p^* do not have a closed form, and are given by

$$p(\lambda_i^* | ELSE, y_i) \propto [e^{p^*} + e^{-\exp(\lambda_i^*)}] I(y_i=0) [e^{-\exp(\lambda_i^*) + \lambda_i^* y_i}] I(y_i>0) \\ \times \exp \left[-\frac{1}{2\sigma_\varepsilon^2} (\lambda_i^* - \mathbf{x}_i' \beta - \mathbf{w}_i' \mathbf{h} - \mathbf{z}_i' \mathbf{s})^2 \right],$$

and

$$p(p^* | ELSE, y_i) \\ \propto [e^{p^*} + e^{-\exp(\lambda_i^*)}] I(y_i=0) [1 + e^{p^*}]^{-1} I(p^*; -5, 5),$$

respectively, where $I(p^*; -5, 5)$ means that $p^* \in (-5, 5)$. Here, $I(\cdot)$ is the indicator function, which is equal to 1 if the condition is satisfied and 0 otherwise; $\mathbf{x}_i'$, $\mathbf{w}_i'$, and $\mathbf{z}_i'$, are the i th rows of $\mathbf{X}_i$, $\mathbf{W}_i$, and $\mathbf{Z}_i$, respectively.

A Gibbs-Metropolis algorithm was developed, based on sampling iteratively from all fully conditional distributions. Location effects, β , $\mathbf{h}$, and $\mathbf{s}$ were sampled from a multivariate normal distribution with appropriate mean and (co)variance matrix, and the 3 variance components were drawn from scale-inverted χ^2 distributions (Wang et al., 1993, 1994). A Metropolis algorithm was used for all other parameters. A normal distribution with mean equal to $\lambda_i^{*(t)}$ at iteration t and an appropriate variance was used as proposal distribution to

sample $\lambda_i^{*(t+1)}$. Similarly, a normal distribution with mean equal to $p^{*(t)}$ at iteration t and an appropriate variance was used as proposal distribution to sample $p^{*(t+1)}$. The variances of the proposal distributions were chosen to satisfy acceptance rates between 30 and 50% (Gelman et al., 2002), which restrict the acceptance rate to being neither too high nor too low, reducing the chance of a chain getting stuck or moving too slowly in the parameter space.

Convergence Diagnostics. Visual inspection of trace plots of the MCMC run and diagnostics suggested by Gelman and Rubin (1992) were used to determine the length of the burn-in period and the total number of iterations for the Gibbs-Metropolis procedure. Two chains with overdispersed starting points were used in the method of Gelman and Rubin (1992), which is based on monitoring convergence of the iterative simulation by estimating the factor by which the scale of the current distribution for a parameter under study, say τ , might be reduced if simulations were continued for an infinite amount of time. The potential scale reduction is given by

$$R = \sqrt{\frac{\text{var}(\tau|\mathbf{y})}{W}},$$

where

$$\text{var}(\tau|\mathbf{y}) = \frac{J}{J-1} W + \frac{1}{J} B.$$

Here, W and B are estimates of the within- and between-chain variances, and J is the number of iterations. R declines to 1 as the number of iterations goes to infinity. There is an indication of convergence when R is close to 1.

Two chains starting from overdispersed values were used, with a total of 500,000 samples, including a burn-in period of 250,000 iterations. A visual inspection of the trace plots (results not shown) suggested that the 2 chains had converged after 250,000 iterations. For the Gelman and Rubin (1992) convergence criterion, the scale reduction factors (R) for the residual, herd, and sire variances were 0.9999996, 1.0001, and 1.000001, respectively; the scale reduction R for p was 1.00001. On this basis, we decided to use only one chain with a burn-in period of 250,000 iterations, and with 250,000 samples collected thereafter for calculating Monte Carlo estimates of posterior features. No thinning of samples was practiced.

Residual Analysis

In the residual analysis, a residual for the record of cow i under a given model was defined as $r_i = y_i -$

Figure 1. Posterior means of the standardized residuals under the zero-inflated Poisson (ZIP) and Poisson models.

$10 - E(y_i|\tau)$. The posterior mean of the standardized residual was estimated as

$$\hat{r}_i = \frac{1}{J} \sum_{j=1}^J \frac{[y_i - E(y_i | \tau^{(j)})]}{\sqrt{\text{var}(y_i | \tau^{(j)})}},$$

with $i = 1, \dots, n$; J is equal to the number of samples from the posterior distribution, and the expectation and variance correspond to those under model M with parameter vector τ . An observation would be unusual if the posterior distribution of r_i were concentrated away from zero. Under the Poisson model, $E(y_i|\tau^{(j)}) = \text{var}(y_i|\tau^{(j)}) = \lambda_i$. Under the ZIP model, $E(y_i|\tau^{(j)}) = (1 - p)\lambda_i$ and $\text{var}(y_i|\tau^{(j)}) = E(y_i|\tau^{(j)})[1 + \lambda_i p]$, where $\lambda_i = \exp(\lambda_i^*)$ and $p = \exp(p^*)/[1 + \exp(p^*)]$.

Posterior Predictive Assessment

The adequacy of a given statistical model may be assessed by referring the observed value of some statistic to its sampling distribution expected under this model (Sorensen and Waagepetersen, 2003). If the ob-

served value is atypical, this would be interpreted as evidence against the model. One way of checking whether a model describes the observed data well is to adopt a measure of discrepancy, say $T(\cdot)$, between the observed data and predictions from the model posited (Gelman and Rubin., 1992).

Let $\mathbf{y}_{rep}$ be “replicated” or simulated data from model M with parameter vector τ , and $\mathbf{y}$ be the observed data vector. In the posterior predictive approach to model validation, one can work with $T(\mathbf{y}|\tau) - T(\mathbf{y}_{rep}|\tau)$ and check whether zero is an extreme value in the posterior predictive distribution of $T(\mathbf{y}|\tau) - T(\mathbf{y}_{rep}|\tau)$. Because the parameter vector τ is unknown, its values are sampled from the posterior distribution. Likewise, $\mathbf{y}_{rep}$ is obtained by sampling from the posterior predictive density

$$p(\mathbf{y}_{rep} | \mathbf{y}, M) = \int p(\mathbf{y}_{rep} | \mathbf{y}, \tau, M) p(\tau | \mathbf{y}, M) d\tau = E_{\tau|\mathbf{y}, M} [p(\mathbf{y}_{rep} | \mathbf{y}, \tau, M)].$$

Here, $p(\mathbf{y}_{rep} | \mathbf{y}, \tau, M) = p(\mathbf{y}_{rep} | \tau, M)$, because $\mathbf{y}_{rep}$ (the new sample, given the parameters) is independent of $\mathbf{y}$ (the observed data, given the parameters).

Figure 2. Posterior realizations of the discrepancy statistic T_1 ($\cdot$, $\tau^{(l)}$) (observed data in abscissa and replicated data in the ordinates) under the zero-inflated Poisson (ZIP) and Poisson models.

Two different discrepancy measures were used, with each of these evaluated at the observed and replicated data. The first measure was based on a sum, across observations, of standardized residuals:

$$T_1(\mathbf{y}, \tau^{(l)}) = \frac{1}{n} \sum_{i=1}^n \frac{[y_i - E_M(y_i | \tau^{(l)})]}{\sqrt{\text{var}_M(y_i | \tau^{(l)})}},$$

with the expectation and variance corresponding to model M . Using the samples from the MCMC procedure, we constructed a scatter plot to measure the discrepancy between $T_1(\mathbf{y}, \tau^{(l)})$ and $T_1(\mathbf{y}_{rep}^{(l)}, \tau^{(l)})$, where $l = 1, \dots, 10,000$ is the number of simulated samples. If the model fits, points are expected to fall in a circle centered at (0, 0).

The second measure of discrepancy used was based on:

$$T_{2,k}(\mathbf{y}) = \frac{\text{number of } k \text{ cases of mastitis}}{n},$$

$k = 0, 1, 2, \dots$, which gives the proportion of cows with k cases of mastitis in a sample of size n . We examined how $D_k^{(l)} = T_{2,k}(\mathbf{y}) - T_{2,k}(\mathbf{y}_{rep}^{(l)})$ varied over k , with $l = 1, \dots, 10,000$ simulated samples, each of size n . Here, $T_{2,k}(\mathbf{y})$ represents the proportion of cows with k cases of mastitis in the observed sample, and $T_{2,k}(\mathbf{y}_{rep}^{(l)})$ represents the proportion of cows with k cases of mastitis in the simulated sample under $\tau = \tau^{(l)}$.

Computations were as follows: 1) each element $\mathbf{y}_{rep,i}^{(l)}$ of the vector $\mathbf{y}_{rep}^{(l)} = (\mathbf{y}_{rep,1}^{(l)}, \dots, \mathbf{y}_{rep,n}^{(l)})$ was drawn from a ZIP distribution evaluated at the posterior sample of λ_i^* and p^* at iteration l ; 2) the statistics $T_1(\mathbf{y}_{rep}^{(l)}, \tau^{(l)})$ and $T_{2,k}(\mathbf{y}_{rep}^{(l)})$ were evaluated for each realization of $\mathbf{y}_{rep}^{(l)}$. This yielded 10,000 “replicated” standardized residuals and relative frequencies of the distribution of mastitis.

Sire Evaluation

The posterior mean of the probability of producing daughters without any case of mastitis in first lactation is appealing for the genetic evaluation and ranking of

Figure 3. Posterior realizations of the discrepancy statistic function $D_k^{(l)}$ under the zero-inflated Poisson (ZIP) and Poisson models.

sires for selection. Sire transmitting abilities in the logarithmic scale for the Poisson parameters do not lead directly to the probability of sires having daughters without mastitis infection, because the Poisson parameter depends on other effects in the model. Let

$$\lambda_j^* = \mu + \bar{h} + s_j$$

be the expected Poisson parameter of sire j , with μ being some combination of levels of the β vector, and $\bar{h}$ being the average of the posterior means of the herd effects. In addition, let P_j be the probability that sire j produces a daughter without mastitis. Then

$$P_j = p + (1 - p)e^{\lambda_j},$$

where $\lambda_j = \exp(\lambda_j^*)$ and $p = \exp(p^*)/[1 + \exp(p^*)]$. Sire ranking could be based on the posterior mean of P_j . However, note that P_j decreases monotonically as s_j increases. Hence, in a model with homogenous p , sire ranking based on posterior means of P_j and s_j are identical, given p .

Correlations between sire evaluations under Poisson and ZIP models were computed by using the posterior

mean of the probability of no mastitis for each of the sires as the criterion. The β -effects were combined by setting calving year to 1992, calving month to August, age at calving to 24 mo, and number of days from first calving to the defined end of the lactation as equal to 300.

RESULTS AND DISCUSSION

Residual plots for the ZIP and Poisson models are shown in Figure 1. The plots indicated that the residuals were larger under the ZIP model than under the Poisson model; however, this difference was very small. The distribution of posterior means of residuals associated with cows having zero or one mastitis case were nearly the same in the 2 models. For counts larger than one, residual posterior means under the ZIP model were slightly larger. Neither model had a completely satisfactory performance in the analysis of residuals.

Figure 2 shows a scatter plot of the statistics $T_1(\mathbf{y}^{(l)}, \boldsymbol{\tau}^{(l)})$ (abscissa) and $T_1(\mathbf{y}_{rep}^{(l)}, \boldsymbol{\tau}^{(l)})$ (ordinate) for the ZIP and Poisson models. Both graphs show that realizations for the observed-data statistic cluster around the coordinate (0, 0), indicating a good agreement between ob-

Figure 4. Posterior distributions of the residual, herd and sire variances, and of the perfect state probability p , under the zero-inflated Poisson (ZIP) model.

served data and predictions, at least with respect to T_1 ($\cdot$, $\tau^{(l)}$).

The discrepancy statistic $D_k^{(l)}$ was used for posterior predictive checking of the ZIP and Poisson models, at each of $k = 0, 1, 2, 3$, and 4 or more cases of clinical mastitis per cow. The posterior predictive distributions of $D_k^{(l)}$ under the ZIP and Poisson models are displayed in Figure 3. The plots suggest that the ZIP model was closer than the Poisson model to the observed proportion of mastitis cases across all number of cases, because the mean values of the discrepancy statistic were closer to zero for ZIP. In particular, the proportions of 0, 1, and 2 cases of mastitis were better predicted by ZIP than by the Poisson model.

Posterior means (standard deviation) of the residual, herd, and sire variances were 0.36 (0.04), 0.37 (0.02), and 0.09 (0.01), respectively, for the ZIP model and 0.90 (0.03), 0.35 (0.02), and 0.05 (0.01) for the Poisson model. In the ZIP model, the most important source of variation was herd, followed by the residual and sire vari-

ances. The herd and sire variances were similar in the ZIP and Poisson models, but the residual variance in the Poisson model was almost 3 times greater than that for the ZIP model, indicating that the residuals in the Poisson model were inflated by the excess of zeros.

In the ZIP model, the herd and sire variances accounted for 37 and 9%, respectively, of the variability of log-Poisson parameters, indicating the relevance of these effects. Posterior distributions of the variance components and of the perfect state probability, p , under the ZIP model are given in Figure 4. The posterior distributions of the residual and herd variances, as well as of the perfect state probability were nearly symmetric. The posterior distribution of the sire variance had a slightly longer tail to the right. This was expected, because the posterior mean of σ_s^2 was closer to 0 than those of the residual and herd components of variance. There is no obviously useful definition of heritability in this model; hence, posterior distributions of ratios of the variance components were not assessed. However,

Figure 5. Distributions of the posterior means of the Poisson parameter ($\lambda_j^* = \mu + \bar{h} + s_j$) given by the “sire j” contribution and of the probability of no mastitis (sire evaluation) under the zero-inflated Poisson (ZIP) model.

the between-sires variance accounted for 10% of the variation in λ_i^* parameters, indicating genetic variation for mastitis.

The posterior distribution of the perfect state probability (Figure 4) was nearly symmetric, and its mean (standard deviation) was 0.32 (0.02). This means that, on average, 32% of first-lactation Norwegian Red cows would be expected to be in the perfect state and not get mastitis at all, either because of complete resistance to the disease, or because they are not exposed to it.

The Monte Carlo variances of residual, herd, and sire variances and of the perfect state probability were 3.3×10^{-8} , 7.5×10^{-9} , 8.6×10^{-8} , and 6.1×10^{-9} , respectively, indicating small Monte Carlo errors.

Figure 5 shows the distribution of the posterior means of sire transmitting abilities under the ZIP model. Under this model, the sire “transmitting ability” e_j^* for the Poisson parameter was 0.31, on average, whereas P_j averaged 0.82, where P_j is the probability that sire j produces a daughter with no mastitis. The correlation between sire ranks based on posterior means of P_j under the Poisson and ZIP models was 0.98.

The rank correlation between the top 10, 20, 30, or 50 out of 437 sires ranged from 0.96 to 0.98. This indicates a strong agreement between sire rankings based on ZIP and Poisson models.

It would be of interest to relax the assumption of a homogenous p . This assumption may hamper the performance of the ZIP model for levels of “fixed” and “random” effects that do not present extra zeros. Varying p across sires, for example, would allow us to separate sires conferring true resistance to mastitis from those that are only mildly liable to the disease.

CONCLUSIONS

The posterior predictive check favored the ZIP specification over the Poisson model. The residual analysis showed very small differences between the ZIP and Poisson models at 0 and 1 cases, but slightly favored the Poisson model for 2 or more cases of mastitis during first lactation. Using the probability of no mastitis as a criterion for genetic evaluation, we found a strong agreement between sire rankings from the Poisson and

ZIP models. A weaker association between sire evaluations is to be expected in a population with stronger inflation at 0, as well as in the situation in which the ZIP model involves a heterogeneous p .

The inferred mixture probability was 0.32, giving a weight of 0.32 to the perfect state and a weight of 0.68 to the imperfect state, represented by a Poisson model. Hence, 32% of first-lactation Norwegian Red cows are expected to be in the perfect state and not contract mastitis at all, either because of complete resistance to the disease or because they are not exposed to it.

ACKNOWLEDGMENTS

Access to the data was given by the Norwegian Dairy Herd Recording System (Ås, Norway) and the Norwegian Cattle Health Service (Ås, Norway) in agreement number 004.2005. Research was funded by grants CAPES BEX 1758004, USDA 2003-35205-12833, NSF DEB-0089742, and NSF DMS 0443771. Support by the Wisconsin Agriculture Experiment Station and by the Babcock Institute for Dairy Research and Development is acknowledged.

REFERENCES

- Barnouin, J., S. Bord, S. Bazin, and M. Chassagne. 2005. Dairy management practices associated with incidence rate of clinical mastitis in low somatic cell score herds in France. *J. Dairy Sci.* 88:3700–3709.
- Breslow, N. E. 1984. Extra-Poisson variation in log-linear models. *Appl. Stat.* 33:38–44.
- Brillinger, D. R. 1986. The natural variability of vital rates and associated statistics (with discussion). *Biometrics* 42:693–734.
- Carlén, E., M. del P. Schneider, and E. Strandberg. 2005. Comparison between linear models and survival analysis for genetic evaluation of clinical mastitis in dairy cattle. *J. Dairy Sci.* 88:797–803.
- Chang, Y. M., D. Gianola, B. Heringstad, and G. Klemetsdal. 2004. Effects of trait definition on genetic parameter estimates and sire evaluation for clinical mastitis threshold models. *Anim. Sci.* 79:355–364.
- Chang, Y. M., D. Gianola, B. Heringstad, and G. Klemetsdal. 2006. A comparison between multivariate slash, Student- t and probit threshold model for analysis of clinical mastitis in first lactation cows. *J. Anim. Breed. Genet.* 123:290–300.
- Damgaard, L. H., I. R. Korsgaard, J. Simonsen, O. Dalsgaard, and A. H. Andersen. 2006. The effect of ignoring individual heterogeneity in Weibull log-normal sire frailty models. *J. Anim. Sci.* 84:1338–1350.
- Foulley, J. L., D. Gianola, and S. Im. 1987. Genetic evaluation of traits distributed as Poisson-binomial with reference to reproductive characters. *Theor. Appl. Genet.* 73:870–877.
- Gasqui, P., O. Pons, and J. Coulon. 2000. An individual modeling tool for consecutive clinical mastitis during the same lactation in dairy cows: A method based on a survival model. *Vet. Res.* 31:583–602.
- Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2002. *Bayesian Data Analysis*. 2nd ed. Chapman and Hall, London, UK.
- Gelman, A., and D. B. Rubin. 1992. Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7:457–511.
- Heringstad, B., Y. M. Chang, D. Gianola, and G. Klemetsdal. 2003b. Genetic analysis of longitudinal trajectory of clinical mastitis in first-lactation Norwegian cattle. *J. Dairy Sci.* 86:2676–2683.
- Heringstad, B., R. Rekaya, D. Gianola, G. Klemetsdal, and K. A. Weigel. 2003a. Genetic change for clinical mastitis in Norwegian cattle: A threshold model analysis. *J. Dairy Sci.* 86:369–375.
- Lambert, D. 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 34:1–14.
- Rao, C. R., and I. M. Chakravarti. 1956. Some small sample tests of significance for a Poisson distribution. *Biometrics* 12:264–282.
- Schukken, Y. H., F. J. Grommers, D. van de Geer, H. N. Erb, and A. Brand. 1990. Risk factors for clinical mastitis in herds with a low bulk milk somatic cell count. Risk factors for *Escherichia coli* and *Staphylococcus aureus*. *J. Dairy Sci.* 73:3436–3471.
- Sorensen, D., and R. Waagepetersen. 2003. Normal linear models with genetically structured residual variance heterogeneity: A case study. *Genet. Res.* 82:207–222.
- Tempelman, R. J., and D. Gianola. 1996. A mixed effects model for overdispersed count data in animal breeding. *Biometrics* 52:265–279.
- van den Broek, J. 1995. A score test for zero inflation in a Poisson distribution. *Biometrics* 51:738–743.
- Wang, C. S., J. J. Rutledge, and D. Gianola. 1993. Marginal inferences about variance components in a mixed linear model using Gibbs sampling. *Genet. Sel. Evol.* 25:41–62.
- Wang, C. S., J. J. Rutledge, and D. Gianola. 1994. Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs. *Genet. Sel. Evol.* 26:91–115.