Analyses of Categorical Data

L. R. Schaeffer

February 19, 2009

1 Introduction

Categorical data have always been a problem for analysis. Most of my work has been with type classification data of dairy cattle, and that is the context for this commentary. The observations on cows are subjective assessments by trained classifiers. The cow is assigned to a 1 to 9 category, or is assigned a number from 60 to 100 (but never 100). The assumption is made that a cow is scored only once in its life, sometime during the first lactation.

In Canada at least, the classifiers gather once or twice a year to be updated on the scoring of cows for the multitude of characteristics that must be scored. The criteria for assigning a cow to a given category for a trait may change due to decisions made by a Type Classification Committee composed of breeders and judges. Changes in criteria are usually changed between rounds. A round is the length of time for classifiers to visit all herds in Canada that are on the program (usually 7 to 9 months). Thus, within a round, all cows should be scored on the basis of the same criteria. Sometimes changes in criteria occur in mid-round, but usually for one or two characteristics, and the change is not very drastic. From one round to the next, however, the criteria can be more thoroughly updated. Thus, a cow that received a score of 80 in the previous round, might only have received a score of 78 in the current round if it had been born a round later.

Classifiers are also given guidelines about the percentages of animals that should receive particular scores. For example, a score above 95 should only happen once in 500 cows (but for first lactation cows, never). Hence, from round to round, even though the criteria for assessment might change, the percentage of animals in each category remains closely the same. This implies that there has been no genetic improvement, because logically if you select future progeny from the better classified animals, then the distribution of progeny classifications should shift towards the higher ("better") end of the scale, but that is not allowed to happen.

Genetic evaluation models currently analyze classifications from many rounds of data (to have suitable numbers of records) as though the same criteria have been used in every round. Sometimes threshold models are used, and sometimes linear models are used. I contend that estimates of genetic trend from these analyses are underestimated, and rankings of animal estimated breeding values (EBVs) are biased. This is supported by problems that CDN (Canadian Dairy Network) has had in validating genetic trends for Interbull comparisons for conformation traits.

2 Proposed Approach

2.1 Data Normalization

Categorical data should be normalized within the time periods in which the criteria of assessing animals has been consistent from animal to animal, that is, within rounds of classification. Determine the percentage of animals with each possible score. Assume an underlying normal distribution, then going from the low end of the scale to the upper end determine the location of the thresholds for each category. Finally, determine the average value of animals between each pair of threshold locations. These averages become the observations to be analyzed. Instead of using 1 for category 1, the normalized value might be -2.33.

By normalizing within rounds, changes in criteria of assessment that might change the percentages of animals in each category are taken into account. Also, the numerical differences between categories are better established than just using category numbers.

CDN has been normalizing the data within rounds for some years now, and this should continue. However, this is not enough to handle the changing criteria from round to round.

2.2 Model of Analysis

Now assume that the data within each round are a different trait, but that genetically the traits are highly correlated. Each animal has only one record in only one round. Thus, there are no environmental correlations between rounds. The ties between rounds are due to progeny of the same sires and same dams being distributed across rounds. The model for data in the k^{th} round can be written as

$$\mathbf{y}_k \;=\; \mathbf{X}_k\mathbf{b}_k + \mathbf{Z}_k\mathbf{h}_k + \mathbf{T}_k\mathbf{a}_k + \mathbf{e}_k,$$

where $\mathbf{y}_k$ are the normalized scores of cows classified in the k^{th} round, $\mathbf{b}_k$ are the fixed effects of classifier within round and age and stage of classification effects within round and $\mathbf{X}_k$ is the associated design matrix for those effects, $\mathbf{h}_k$ are random contemporary group effects (herd-round-classifier subclasses), and $\mathbf{Z}_k$ is the design matrix for contemporary groups, $\mathbf{a}_k$ are the additive genetic effects of trait k for all animals in the data with $\mathbf{T}_k$ indicating the animals with records in round k, and $\mathbf{e}_k$ are the random residual effects of animals classified in round k.

Assume that

$$\mathbf{R} \;=\; Var\begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \vdots \\ \mathbf{e}_r \end{pmatrix} \;=\; \begin{pmatrix} \mathbf{I}\sigma_e^2 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\sigma_e^2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}\sigma_e^2 \end{pmatrix},$$

$$\mathbf{H} \;=\; Var\begin{pmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \\ \vdots \\ \mathbf{h}_r \end{pmatrix} \;=\; \begin{pmatrix} \mathbf{I}\sigma_h^2 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\sigma_h^2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}\sigma_h^2 \end{pmatrix},$$

and

$$\mathbf{G} \;=\; Var\begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_r \end{pmatrix} \;=\; \begin{pmatrix} \mathbf{A}g_{11} & \mathbf{A}g_{12} & \cdots & \mathbf{A}g_{1r} \\ \mathbf{A}g_{12} & \mathbf{A}g_{22} & \cdots & \mathbf{A}g_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}g_{1r} & \mathbf{A}g_{2r} & \cdots & \mathbf{A}g_{rr} \end{pmatrix}.$$

Let

$$\mathbf{W} = \begin{pmatrix} g_{11} & g_{12} & \cdots & g_{1r} \\ g_{12} & g_{22} & \cdots & g_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ g_{1r} & g_{2r} & \cdots & g_{rr} \end{pmatrix},$$

then

$$\mathbf{G} = \mathbf{A} \bigotimes \mathbf{W},$$

and

$$\mathbf{G}^{-1} \;=\; \mathbf{A}^{-1} \bigotimes \mathbf{W}^{-1}.$$

2.3 Simplifications

Because fixed effects and random contemporary group effects are nested within rounds, then the model can be viewed as one single trait model, but with multiple values for the animal additive genetic effects, one for each round. If σ_e^2 and σ_h^2 are the same for each round, then the mixed model equations are simple to construct (as scalars during the iteration on data process). The only tricky part is for the animal genetic effects. Let A_{ii} be the diagonal element of $\mathbf{A}^{-1}$ for animal i, and let the genetic covariances in $\mathbf{W}$ be all the same, and the genetic variance of each round is the same, then

$$\mathbf{W} \;=\; \begin{pmatrix} 1 & c & \cdots & c \\ c & 1 & \cdots & c \\ \vdots & \vdots & \ddots & \vdots \\ c & c & \cdots & 1 \end{pmatrix} \sigma_g^2 \;=\; (\mathbf{I}x + \mathbf{J}z)\sigma_g^2,$$

where $z = c$, the genetic correlation between rounds, and $x = 1 - c$. The inverse of $\mathbf{W}$ has the same structure and depends on x, z, and the order of $\mathbf{W}$, i.e. number of rounds.

During the iteration on data process, the vector of r solutions for animal i is given by

$$\hat{\mathbf{a}}_i = (\mathbf{Q} + A_{ii}\mathbf{W}^{-1}\alpha)^{-1}\mathbf{p},$$

where $\mathbf{Q}$ is a null, square matrix with a single 1 in it, located on the diagonal of the round in which the animal was classified, α is the ratio of residual variance to additive genetic variance, and $\mathbf{p}$ is a vector that contains $-\mathbf{W}^{-1}\mathbf{A}^{-1}\alpha\hat{\mathbf{a}}_{-i}$, (adjustments for parents, progeny, and mates of animal i) plus

$$(y_{ik} - \hat{RC} - \hat{AS}R - \hat{h}),$$

the cow's classification adjusted for the current values of the round-classifier solutions, age-stage of classification solutions and contemporary group solution, added to the k^{th} element of $\mathbf{p}$ corresponding to the round in which the cow was classified.

Interestingly,

$$(\mathbf{Q} + A_{ii}\mathbf{W}^{-1}\alpha)^{-1} = \mathbf{C}$$

has an unique structure too, given the previous assumptions. In general, let k be the round in which the cow was classified, then

$$\mathbf{C} = \begin{pmatrix} c_1 & c_2 & \cdots & c_3 & \cdots & c_2 \\ c_2 & c_1 & \cdots & c_3 & \cdots & c_2 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ c_3 & c_3 & \cdots & c_4 & \cdots & c_3 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ c_2 & c_2 & \cdots & c_3 & \cdots & c_1 \end{pmatrix}.$$

Thus, there are only 4 different numbers in $\mathbf{C}$, which depend on the two elements of $\mathbf{W}^{-1}$, the round, and the number of rounds or order of $\mathbf{C}$. So if there are 50 rounds to be analyzed, the $\mathbf{C}$ matrix can be represented completely by just 4 numbers and k, the round in which the record occurred. Normally, $\mathbf{C}$ would have 1275 elements if 50 rounds were involved, but only 4 different values. These can be easily calculated in one pass through the data. If the animal is a sire (no classification record), then $\mathbf{C}$ has just 2 different values similar to the structure of $\mathbf{W}$.

Thus, with only a little modification, the current genetic evaluation programs can be made to handle multiple trait models.

2.4 Estimated Breeding Values

Every animal receives r estimated breeding values, i.e. one for each round, even though the animal was classified in only one round. By using a high genetic correlation, like 0.98, between rounds, the solutions for the different rounds are nearly identical (not exactly, but nearly). These can be averaged to deliver just one EBV per animal.

The normalization of data, the multiple trait model, and the high genetic correlation all combine to give EBV that are more highly correlated to the animal's true breeding values. Consequently, genetic trends will not be biased, and should pass the validation tests for use in Interbull. To me, this is the best way to analyze type classification data, or categorical data similar to it.

3 Conclusions

There have been reports on this work presented to the Dairy Cattle Breeding and Genetics Committee in February 2009, and a final report in September 2009 will be made on this method applied to actual data.

This method would not be appropriate for all categorical data situations, such as health traits, where the percentage of animals in the diseased category, for example, can decrease over time. That means the criteria for assessing diseased or not diseased remains the same, and there is no question about which category an animal belongs. The population is shifting in the proportions of diseased and healthy animals. In this case, a threshold model may be better to use.

Otherwise, you should normalize your data within time periods, and apply a multiple trait model where data in each time period is a separate trait. Application was easy in the type classification example because cows were only observed once and only in one round. If animals were observed more than once and in different time periods, then a multiple trait application may be slightly more complex to implement because there would be residual correlations between observations on the same animal.