Title: **Necessary Changes to Improve Animal Models**

L. R. Schaeffer

Department of Animal Biosciences, University of Guelph, Guelph, Ontario, N1G 2W1, Canada

Running Title: **Animal Models**

Summary

Animal models evolved from sire models and inherited some issues that affected sire models. Those include definition and treatment of contemporary groups, accounting for time trends, and dealing with animals having unknown parents. The assumptions and limitations of the animal model need to be kept in mind. This review of the animal model will discuss the issues and will recommend enhancements to animal models for future applications.

Keywords: Animal model
Assumptions
Limitations
Phantom Parent Groups
Contemporary Groups

Introduction

In 1970, the animal breeding world was introduced to linear models and best linear unbiased prediction methods (BLUP) by C. R. Henderson through the Northeast AI Sire Comparison. The initial model was

$$y_{ijklm} \;=\; YS_i + HYS_{ij} + G_k + S_{kl} + e_{ijklm},$$

where

y_{ijklm} was first lactation 305-d milk yield of daughter m of sire l belonging to genetic group k making a record in year-season of calving i and herd-year-season j;

YS_i was a fixed year-season of calving effect to account for time trends in the data;

HYS_{ij} was a random herd-year-season of first calving contemporary group;

G_k was a fixed sire genetic group, defined by the year of sampling and AI ownership;

S_{kl} was a random sire effect within genetic group; and

e_{ijklm} was a random residual effect.

In matrix notation, let

y be the vector of first lactation milk yields,

b be the vector of year-season effects,

h be the vector of herd-year-season effects,

g be the vector of genetic group effects,

s be the vector of sire transmitting abilities, and

e be the vector of residuals,

then

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Wh} + \mathbf{Qg} + \mathbf{Zs} + \mathbf{e},$$

where $\mathbf{X}$, $\mathbf{W}$, $\mathbf{Q}$, and $\mathbf{Z}$ are design matrices relating observations to the factors in the model.

Also,

$$\begin{aligned} E(\mathbf{y}) &= \mathbf{Xb} + \mathbf{Qg} \\ E(\mathbf{h}) &= \mathbf{0} \\ E(\mathbf{s}) &= \mathbf{0} \\ E(\mathbf{e}) &= \mathbf{0} \\ Var(\mathbf{e}) &= \mathbf{I}\sigma_e^2 \\ Var(\mathbf{s}) &= \mathbf{I}\sigma_s^2 \\ Var(\mathbf{h}) &= \mathbf{I}\sigma_h^2 \end{aligned}$$

The assumptions of this model were

1. Sires were unrelated to each other.
2. Sires were mated randomly to dams.
3. Progeny were a random sample of daughters.
4. Daughters of sires were randomly distributed across herd-year-seasons.
5. Milk yields were adjusted perfectly for age and month of calving.

The limitations were

1. Sires were related to each other.
2. Because they were AI sires, semen prices varied depending on the background of the bull. Sires were not randomly mated to dams in the population.

3. Daughters of higher priced bulls tended to be associated with richer herds that supposedly had better environments.

4. The age-month of calving adjustment factors were not without errors.

5. Only first lactation records were used.

6. Cows were not evaluated.

Selection Bias

At the time, the industry believed the existence of a non-random association of the true values of sires and herd-year-seasons. Henderson (1975) had a theory about different kinds of selection bias and how to account for them. Henderson outlined three types of selection. Selection on $\mathbf{y}$, or phenotypic selection; Selection on $\mathbf{u}$, a random factor in the model; and Selection on $\mathbf{e}$, or affecting the residual variation associated with animal performance. Henderson assumed that a matrix, $\mathbf{L}'$, existed such that it described the difference between selected and non-selected elements. Unfortunately, Henderson did not give any general instructions on how $\mathbf{L}'$ was to be constructed or what it might look like, only that such a matrix existed. Some examples are shown in his book (Henderson, 1984). Each type of selection resulted in a different set of modified mixed model equations which included the $\mathbf{L}'$ matrix.

For the sire by herd association bias, Henderson's solution was to treat either sire effects or herd-year-season effects as a fixed factor in the model, in which case the modified MME gave the correct expectations of the random sire effects under the selection model. There was no need to construct $\mathbf{L}'$ explicitly. Henderson chose to make herd-year-seasons fixed, although he could have as easily made sires fixed instead. Think what that would have done to genetic evaluations.

Thompson (1979) argued that $\mathbf{L}'$ was not well defined, and was arbitrarily random. Gianola et al. (1988) argued against the concept of repeated sampling underlying the assumptions of Henderson's theory. Note that making herd-year-seasons fixed to remove bias only works for the sire model with random sire and random contemporary groups, and only if you believe Henderson's theory of selection is correct.

Thus, Henderson modified the initial model into

$$\mathbf{y} = \mathbf{Wh} + \mathbf{Qg} + \mathbf{Zs} + \mathbf{e},$$

where $\mathbf{W}$, $\mathbf{Q}$, and $\mathbf{Z}$ are design matrices relating observations to the factors in the model. Note that $\mathbf{h}$ are now fixed effects in the model and that year-seasons were totally confounded with herd-year-seasons. Consequently, $\widehat{\mathbf{s}}$ were not biased by any association of sires with herd-year-seasons in this modified version, according to Henderson's theory.

Also,

$$\begin{aligned}
E(\mathbf{y}) &= \mathbf{Wh} + \mathbf{Qg} \\
E(\mathbf{s}) &= \mathbf{0} \\
E(\mathbf{e}) &= \mathbf{0} \\
Var(\mathbf{e}) &= \mathbf{I}\sigma_e^2 \\
Var(\mathbf{s}) &= \mathbf{I}\sigma_s^2
\end{aligned}$$

In the northeast United States, contemporary groups were fairly large for each herd-year-season, but in some European countries there were many contemporary groups with fewer than five animals. Any contemporary groups with all daughters from only one bull did not contribute any information to sire evaluations.

The other assumptions and limitations were as with the initial model. The problems of sires being related, and not being randomly mated to dams were probably much more significant in their effects on estimated transmitting abilities than the problem of non-random association of sires with herd-year-season effects, but were largely ignored.

The modified model is the one that every country tried to adopt during the 1970's, and with my help. Thus, it became common practice to have fixed contemporary groups in sire models, even if the bias that was present in the northeast United States did not exist in other countries or situations. For example, sire models used in swine or sheep, where artificial insemination was not prevalent and where progeny group sizes were not large, probably had no selection bias needing removal.

Sires Related

Henderson (1976) discovered a method of inverting the additive genetic relationship matrix ($\mathbf{A}$), and this made it possible to account for sires that were related, through their sire and maternal grandsire. Herd-year-seasons were still treated as a fixed factor. The model did not account for non-random mating of sires to dams. Now

$$Var(\mathbf{s}) = \mathbf{A}\sigma_s^2.$$

The sire model continued to be employed for sire evaluation until 1988. By 1988, computer hardware and computing techniques had improved to make animal models feasible (Meyer and Burnside, 1988).

In summary, the problems with sire models were

- Sires not randomly mated to dams.
- Having enough bulls in each genetic group.
- Only first lactations of cows were used.
- No cow evaluations were produced.
- HYS with all cows being daughters of the same bull were useless.

Problems motivate changes for the future, and the problems of the sire model motivated change to an animal model.

Animal Models

Papers by Thompson (1979) and Gianola et al. (1988) criticized the selection bias theories of Henderson (1975) and effectively stopped future discussion about them. The fact that $\mathbf{L}'$ selection theory was deemed incorrect, meant that the sire model with herd-year-seasons as fixed effects might not be appropriate, but everyone around the world still used fixed contemporary group effects. Few people understood Henderson's selection bias theories, but if Henderson treated contemporary groups as fixed, then so would they. Even in 2017, contemporary groups are frequently modelled as a fixed factor for animal models, which goes against Henderson's own derivation for a specific sire model.

Several papers have argued about having fixed or random contemporary groups (Ugarte et al., 1992; Van Vleck, 1987; and Visscher and Goddard, 1993). If contemporary group size is large (e.g., 20 or more individuals), then there is little difference in analyses if they are a fixed or random factor. However, some of these studies were not correct. When contemporary groups, $\mathbf{Wh}$, are made random in a model, then it becomes necessary to add $\mathbf{Xb}$, year-season phenotypic time trends, as a fixed factor, back into the model otherwise biases can occur. Genetic trends are estimated through the $\mathbf{A}$ matrix as long as pedigrees trace back to the base generation and all data are included.

Contemporary Groups

Contemporary groups (CG) should always be a random factor in animal models and sire models. CG identify a group of animals that are

- The same gender,
- Born in the same year-month,
- Raised in the same herd, pen, cage, barn, or field,
- Receiving the same feed and management care, and
- Undergoing the same environmental conditions together.

As animals grow and enter different phases of their life, they may change to different contemporary groups. Instead of being born in the same year-month, they may be born in the same year-season (a group of three or four months combined).

The number of animals going into a CG is not known ahead of time, but are formed as events unfold. Putting together a group of animals of the same age and gender is a completely random event, and the effect of that grouping on the animals in the group is randomly generated. By their definition and manner of creation, CG are a random factor for any linear model analysis (LaMotte, 1983).

The Model

There are a few elements which should be present in all animal models. They are

- Fixed time period effects,
- Random contemporary group effects,
- Random animal additive genetic effects, and
- Phantom parent groups because there are always animals with unknown parents.

The equation of the model is written as

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Wu} + \begin{pmatrix} \mathbf{Z} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{a}_w \\ \mathbf{a}_o \end{pmatrix} + \mathbf{Zp} + \mathbf{e}$$

where

- $\mathbf{b}$ is a vector of fixed effects (such as age, year, gender, farm, cage, diet) that affect the trait of interest, and are not genetic in origin,
- $\mathbf{u}$ is a vector of random factors (such as contemporary groups and others),
- $\mathbf{p}$ is a vector of permanent environmental effects,
- $\mathbf{a}_w$ are animals with records, and $\mathbf{a}_o$ are animals without records in $\mathbf{y}$, but which are related to animals in $\mathbf{a}_w$,
- $\mathbf{e}$ is a vector of residual errors,
- $\mathbf{X}$, $\mathbf{W}$, and $\mathbf{Z}$ are desgin matrices relating observations in $\mathbf{y}$ to factors in the model.

Also,

$$\begin{aligned} E(\mathbf{b}) &= \mathbf{b} \\ E(\mathbf{u}) &= \mathbf{0} \\ E\begin{pmatrix} \mathbf{a}_w \\ \mathbf{a}_o \end{pmatrix} &= \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix} \end{aligned}$$

$$E(\mathbf{p}) = \mathbf{0}$$
$$E(\mathbf{e}) = \mathbf{0}$$
$$Var\begin{pmatrix} \mathbf{u} \\ \mathbf{a}_w \\ \mathbf{a}_o \\ \mathbf{p} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \mathbf{U} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{ww}\sigma_a^2 & \mathbf{A}_{wo}\sigma_a^2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{ow}\sigma_a^2 & \mathbf{A}_{oo}\sigma_a^2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_p^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{R}\sigma_e^2 \end{pmatrix}.$$

where

$$\mathbf{U} = \Sigma_i^+ \mathbf{I}\sigma_i^2,$$

for i going from 1 to the number of other random factors in the model, and $\mathbf{R}$ is usually diagonal, but there can be different values on the diagonals depending on the situation. Often all of the diagonals are the same, so that $\mathbf{R} = \mathbf{I}\sigma_e^2$.

The additive genetic relationship matrix is

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{ww} & \mathbf{A}_{wo} \\ \mathbf{A}_{ow} & \mathbf{A}_{oo} \end{pmatrix}.$$

Some of the assumptions of an animal model are

1. Random factors follow normal distributions.
2. Progeny of sire-dam pairs are random from amongst all possible progeny of that pair.
3. Selective matings of sires to dams are taken into account through the relationship matrix (Kennedy et al. 1988).
4. Animals are randomly dispersed across levels of fixed factors.
5. Observations are taken on either males or females, but if taken on both sexes, then the assumption is that parents would rank the same if based only on one gender or the other.
6. No preferential treatment has been given to individuals or groups of individuals.

7. Data should not be a selected subset of all possible animals.

8. Every animal is able to express their full genetic potential without restraint from other individuals within their contemporary groups.

9. Animals can be traced to a common base population of unselected and randomly mating individuals.

There should never be a need to pre-adjust observations for any factor. With today's hardware and software, these factors can be placed in the animal model and solved simultaneously with the other factors of the model. Such factors may interact with time. For example, in dairy cattle, differences between age groups or months of calving can change over the years. An interaction of age and month of calving with years of calving (five year periods maybe) is needed in the animal model. Suppose in 1973 the difference between 24 months of age and 30 months of age was 200 kg of milk. In 2010 the difference between 24 and 30 month old heifers might be 250 kg. Any fixed factor in any animal model may need an interaction with time. Models should be considered to be dynamic and constantly evolving.

Besides time trends, these trends may be localized to different areas of a country. A mountainous country may see differences due to altitudes of the farms. A large country, like Canada, may see differences between west coast, east coast, the eastern provinces and the western provinces. Thus, time trends within regions would be warranted in a national animal model.

Phantom Parent Groups

Animals should have both parents identified as much as possible. The onus should be on the herd owners to provide that information, and on the recording organization to verify the information. Identity tags are prevalent in the livestock industries now to monitor movement of animals within and between countries for health reasons. Even so, individuals creep into the system with unknown parents. Phantom parent groups (Quaas 1988) was an excellent solution. Groups are based on country, population, or breed, and within those follow the four pathways of selection, namely, Sires of Males, Dams of Males, Sires of Females, and Dams of Females. Then within the pathways, year of birth of the offspring. In most species there is unequal selection intensity on each pathway. As time goes by, the genetic level of each pathway changes at different rates.

In Quaas (1988) phantom parent groups were an additional fixed factor in the model. As such, there were often estimability problems because the male and female pathways were often very similar for a given year of birth. Even by changing the composition of groups between male and female pathways so they were not completely confounded, there remained estimability problems with other fixed factors. Thus, the practice of adding one times the variances ratio (σ_e^2 to σ_a^2) to the diagonals of the phantom parent group equations in the mixed model equations began. The dependencies are removed, and the estimated breeding values tend to look normal. Phantom parent groups are simply treated as another animal, but with unknown parents, and the rules of Henderson (1976) for creating $\mathbf{A}^{-1}$ are followed, as shown by Quaas (1988). Implementation of phantom parent groups is relatively simple following Quaas (1988).

Random Regression Model

One type of animal model is a random regression model. Most applications of test day models using random regressions, have a scalar factor for fixed herd-test-date subclasses as a contemporary group. This is incorrect. Herd-test-date groups contain cows that have calved at different times of the year, are at a different point in their lactation, may not have been managed in the same manner, and may even be different lactations. The only thing they share is that they were measured for their test day yields on the same day in a herd. This seems to be totally contrary to the definition of a contemporary group. A herd-test-date group is a very heterogeneous composition of herd mates. I was unfortunately the person who propogated this factor into test day models, but I hereby acknowledge that I was totally incorrect in publishing it.

A better solution is to have a random regression model with parity-year-month of calving fixed curves to account for shifts in curves over time, and herd-parity-year-season of calving random curves as contemporary groups. Thus, the contemporaries would be the same age, same stage of lactation, and same management group. This would be a more homogeneous grouping, and the cows within this group would also tend to be tested at the same time. The rest of the model would be the same as originally published, except that herd-test-date subclasses would be removed.

The fixed curves should attempt to follow the phenotypic curves as closely as possible. A fourth order Legendre polynomial function may not be suitable for this purpose. A spline function may be necessary, or making days in milk periods of 5 or 10 days each throughout the lactation period (i.e. 36 ten day periods, or 72 five day periods), which would model the phenotypic shape almost precisely. The random regressions model the fluctuations around the phenotypic curves.

Conclusions

Considering the definition of contemporary groups, and the fact that everyone now uses an animal model for data analyses, then contemporary groups should **always** be a random factor in the model. Time trends should always be in the model, and phantom parent groups should be used most of the time.

Fixed contemporary groups are a carry-over tradition from sire models where Henderson said that contemporary groups should be fixed in order to avoid the bias of better sires being associated with the better contemporary groups (herd effects), which was based on a theory that subsequent scientists have criticized severely.

For random regression test day models for dairy cattle, contemporary groups should not be defined as herd-test-date subclasses. A better contemporary group for such models would be parity-herd-year-seasons of calving subclasses. The fixed curves should be based on regions-years-months of calving, and should be modelled by either spline functions or by days-in-milk categories.

Changing everyone's minds at this stage in history will be difficult, if not impossible, but I hope some will heed these words.

References

Gianola, D. , Im, S., Fernando, R. L. 1988. Prediction of breeding values under Henderson's selection model: A revisitation. J. Dairy Sci. 71:2790-2798.

Henderson, C. R. 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423-448.

Henderson, C. R. 1976. A simple method for computing the inverse of a numerator relationship matrix used for prediction of breeding values. Biometrics, 32:69-74.

Henderson, C. R. 1984. Applications of Linear Models in Animal Breeding. University of Guelph.

Kennedy, B. W. , Schaeffer, L. R., Sorensen, D. A. 1988. Genetic properties of animal models. J. Dairy Sci. 71:17-26.

LaMotte, L. R. 1983. Fixed-, random-, and mixed-effects models. In Encyclopedia of Statistical Sciences (S. Kotz, N. L. Johnson and C. B. Read, eds.) 3: 137141.

Meyer, K. , Burnside, E. B., 1988. Joint sire and cow evaluation for conformation traits using an individual animal model. J. Dairy Sci. 71:1034-1049.

Quaas, R. L. , Pollak, E. J. 1980. Mixed model methodology for farm and ranch beef cattle testing programs. J. Anim. Sci. 51:1277.

Quaas, R. L. , Pollak, E. J. 1981. Modified equations for sire models with groups. J. Dairy Sci. 64:1868-1872.

Quaas, R. L. 1988.Additive genetic model with groups and relationships. J. Dairy Sci. 71:1310-1318.

Thompson, R. 1979. Sire evaluation. Biometrics 35:339-353.

Ugarte, E. , Alenda, R., Carabano, M. J. 1992. Fixed or random contemporary groups in genetic evaluations. J. Dairy Sci. 75:269-278.

Van Vleck, L. D. 1987. Contemporary groups for genetic evaluations. J. Dairy Sci. 70:2456-2464.

Visscher, P. M. , Goddard, M. E. 1993. Fixed and random contemporary groups. J. Dairy Sci. 76:1444-1454.