USE OF FRACTIONAL POLYNOMIAL FOR MODELLING SOMATIC CELL SCORES IN DAIRY CATTLE

C. Robert-Granié[1], E. Maza[1], R. Rupp[1] and JL. Foulley[2]

[1]INRA-SAGA, BP 27, 31326 Castanet-Tolosan, [2]INRA-SGQA, 78350 Jouy-en-Josas, France

INTRODUCTION

Somatic cell count (SCC) has been widely advocated as an indicator trait for mastitis. In many countries, SCC is measured on a large scale in national milk recording systems (usually at a monthly basis) and used as an indirect criterion to achieve selection for mastitis resistance (Mrode and Swanson, 1996). Genetic models for SCC are mostly based on lactation average of somatic cell scores (SCS = log transformed SCC to achieve normality of distribution). Use of models for test day observations (Reents *et al*., 1995) should, however, better account for short term environmental variation and allow to use all information without restriction on the number of records available or length of time intervals. Therefore, genetic evaluation of this trait should be improved. To that respect, test day models may also account more precisely for short time variation of SCC than lactation models and be more efficient to predict clinical cases and infections in general.

The purpose of this study was to investigate implementation of test day models for SCS. At this stage our interest was restricted to modelling the average trend of SCS during the lactation. Since many observations are available per animal (about 40), we considered a precise polynomial fit of the data and we used to that purpose the technique of fractional polynomials (Royston and Altman, 1994).

MATERIAL AND METHODS

The study was based on a survey conducted in an INRA farm (Le Pin au Haras) between 1998 and 1999 which overall aim was to assess the relationships between somatic cell count and mastitis. Data from 104 primiparous F1 cows (Hosltein x Normande crossbreed) were analysed. Available information consisted of 4,231 SCS, e.g. one record per week and per cow over a period from d5 to d305 after calving. Therefore, about 40 records per individual were used for modelling the lactation curve of SCS.

Conventional polynomials (with positive integer powers) are a classical choice for modelling the relationship between the response variable and one or several continuous covariates. However the curve does not usually fit the data well both at the low and high orders of those polynomials. At low orders, there is little choice among the curve shapes. At high orders, the fit is usually bad at the extremes showing the usual waviness and end-effects. Several techniques are available to fit more acceptable models. Among those, we have chosen the technique of fractional polynomials due to its simplicity, flexibility and parsimony. Fractional polynomials are an extension of conventional polynomials but with real powers. They often appear as an "ad hoc" procedure in the applied statistics literature.

Let t be a positive real covariable, $\mathbf{p} = \{ p_j \}$; $j = 0,1,...,m$ a $(m+1)$ vector of ordered powers so-called the vector of degree m and $\boldsymbol{\xi} = \{ \xi_j \}$ the vector of the corresponding real

coefficients. A fractional polynomial of degree m is defined as follows :
$\phi_m(t,\xi:\mathbf{p}) = \sum_{j=0}^{m} \xi_j \mathrm{H}_j(t)$ where $\mathrm{H}_j(t) = t^{(p_j)}$ if $p_j \neq p_{j-1}$ and $\mathrm{H}_j(t) = \mathrm{H}_{j-1}(t)\ln(t)$ if $p_j = p_{j-1}$. In this last formula $t^{(p_j)}$ represents the Box-Tidwell transformation i.e., $t^{(p_j)} = t^{p_j}$ if $p_j \neq 0$ and $t^{(p_j)} = \ln t$ if $p_j = 0$. At the origin, $\mathrm{H}_0(t) = 1$. For example, for $\mathbf{p} = (0,0,1)$, $\phi_3(t;\xi,\mathbf{p}) = \xi_0 + \xi_1 \ln t + \xi_2 (\ln t)^2 + \xi_3 t$.
If elements of $\mathbf{p}$ are continuous, $\phi_m(t,\xi:\mathbf{p})$ is a non-linear model with parameters $(\xi,\mathbf{p})$. Then, the quantity $D(m,\mathbf{p}) - D(m,\hat{\mathbf{p}})$ where $D = -2\text{loglikelihood}$ and $\hat{\mathbf{p}}$, the MLE of $\mathbf{p}$, has an asymptotic chi-square distribution with m degree of freedom. In practice, $\mathbf{p}$ is restricted to m values selected from a fixed set P of powers (usually fractions but not necessarily). Here we took P={-2, -1, -1/2, -1/3, 0, 1/3, 1/2, …, max(3, m)}. Let $\tilde{\mathbf{p}}$ the p-value associated with the model of lowest deviance, its deviance $D(m,\mathbf{p})$ is larger than $D(m,\tilde{\mathbf{p}})$ so that $D(m,\mathbf{p}) - D(m,\tilde{\mathbf{p}})$ can be viewed as a conservative test for a given value of $\mathbf{p}$. That is why Royston and Altman (1986) proposed to select models for which the difference is lower or equal to an m df chi-square value exceeded with a probability of 0.9. Usually results are presented as a "gain" G which corresponds to the decrease in deviance from a straight line model : $G = G(m,\mathbf{p}) = D(1,1) - D(m,\mathbf{p})$. The model for other fixed effects (calving year, calving season, age at first calving) and for the interaction between them and with transformed covariates was selected using the robust procedure of testing fixed effects proposed by Liang and Zeger (1986) and described by Robert-Granié and Foulley (2001).

RESULTS AND DISCUSSION

The analysis was restricted to the first two degrees (m=1,2). Results are shown on figure 1. Gain is plotted against p_1 values with different curves for p_2 values. The best p and sub optimal p values models are shown in table 1. Optimum p values are $p_1 = p_2 = -1/3$ (i.e., $\xi_0 + \xi_1 t^{-1/3} + \xi_2 t^{-1/3} \ln t$) with a gain of 68.4. However, there are several combinations of p_1 and p_2 values which lead to gains very close to the maximum one, thus allowing some flexibility in the choice of the final model. The p_1 and p_2 values found here were observed in good agreement with a traditional non linear fit of the data for which $\hat{p}_1 = \hat{p}_2 = -0.31$.

As a matter of fact, the Ali-Schaeffer function (1987) which involves t^*, t^{*2}, $\ln t^*$ and $(\ln t^*)^2$ with time variable $t^* = t/305$, can also be interpreted as a fractional polynomial of degree 4 with p-values of 0, 0, 1, 2. The model leads to a deviance of $D = 15922$ and a gain of $G = 69.6$ slightly better (but not significantly) than the one obtained with the best second degree polynomial previously selected. Moreover, the second degree polynomial involves less parameters and fits seemingly better the right part of the curve at the end of lactation (see figure 2). However, more work is needed to extend the comparison among models with different p values for higher degrees (3, 4, etc…)

For a given mean model, there are different competing variance covariance structures which can be compared. Table 2 gives an example of such a comparison involving a conventional second degree polynomial, the second fractional polynomial $p_1 = p_2 = -1/3$ and the Ali-Schaeffer function.

Table 1. Gain for different p values

p_1	p_2	Deviance	Gain
-1	0	15927.17	64.4
-1	1/3	15923.82	67.8
-1	1/2	15924.26	67.3
-1/2	-1/2	15927.37	64.2
-1/2	-1/3	15924.46	67.1
-1/2	0	15923.53	68.1
-1/2	1/3	15927.88	63.7
-1/3	**-1/3**	**15923.21**	**68.4**
-1/3	0	15925.21	66.4

Table 2. Mean and variance models

Fixed	Random	-2RL	-2AIC	-2BIC
F	F	10921	10935	10953
F	AS			
F	C	10859	10873	10891

F : Fractional polynomial (-1/3,-1/3),
AS : Ali-Schaeffer function
C : Conventional second degree polynomial
RL, AIC, BIC : Restricted likelihood, Akaike and Schwartz criteria

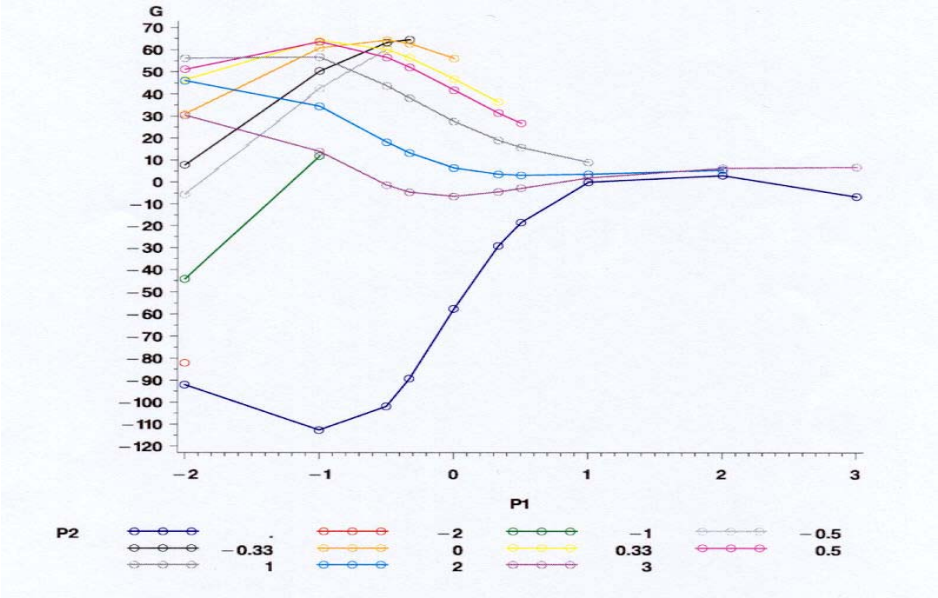

Figure 1. Gain

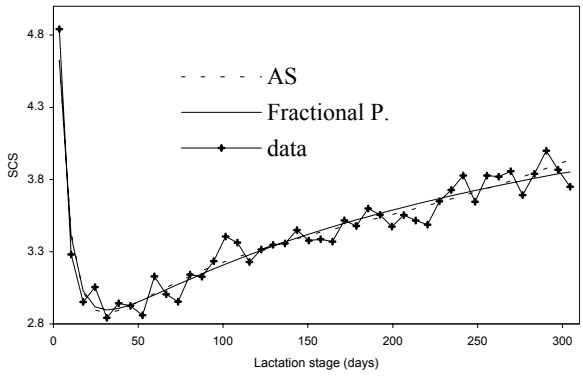

Figure 2. Fit of the mean profile by different functions

The model based on the latter does not converge to a definitive positive variance covariance matrix due probably to a too high degree adjustment. Among the two remaining ones, a conventional second degree random coefficient model appears to be better than the fractional polynomial with $p_1 = p_2 = -1/3$. This example clearly shows that functions selected at the expectation level are not necessarily adequate for the covariance level. At this level, a further distinction could also be made between genetic and permanent environmental effects.

CONCLUSION

As far as the model for means is concerned, fractional polynomials turn out to be a flexible and easy to implement technique as compared to alternative ones (e.g., cubic splines). In particular, this ability was clearly illustrated in the case of the mean profile of SCS during the lactation, the pattern of which remains quite complicated. We were able to fit this mean profile with just a second degree polynomial, thus indicating how parsimonious this procedure can be.

A general study must be undertaken both at the mean and the variance-covariance levels to select the appropriate degrees and p values of the polynomial adjustments to use for these two levels. The choice of the final joint models is not an easy one as there is a strong dependency between the mean and the covariance structures. Eventually, this choice should be based not only on usual model comparison criteria (Deviance, Akaike, Schwartz) but also in relation to the final objective for adjusting SCS (indicator of mastitis) and their genetic variation and interpretation.

REFERENCES

Ali, T.E. and Schaeffer, L.R. (1987) *Can. J. Anim. Sci*. **67** : 637-644.
Liang, K.Y. and Zeger, S.L. (1986) *Biometrika* **73** : 13-22.
Mrode R.A. and Swanson G.J.T. (1996) *Anim. Breed. Abs*. **64** : 847-857.
Reents, R. and Dekkers, J.C.M. (1995) *J. Dairy Sci.* **78** : 2858-2870.
Robert-Granié, C. and Foulley, J.L. (2001) *Proc XXXIII Journées de Statistiques, Nantes*, 657.
Royston, P. and Altman, D.G. (1994) *Appl. Stat*. **43** : 429-467.