Anim. Res. 53 (2004) 259–273
© INRA, EDP Sciences, 2004
DOI: 10.1051/animres:2004016

Original article

Statistical analysis of somatic cell scores via mixed model methodology for longitudinal data

Christèle ROBERT-GRANIÉ[a]*, Jean-Louis FOULLEY[b], Elie MAZA[a], Rachel RUPP[a]

[a] Station d'Amélioration Génétique des Animaux, Institut National de la Recherche Agronomique, BP 27, 31326 Castanet-Tolosan, France
[b] Station de Génétique Quantitative et Appliquée, Institut National de la Recherche Agronomique, 78352 Jouy-en-Josas Cedex, France

(Received 28 November 2002; accepted 29 April 2004)

Abstract – The aim of this study was to analyze somatic cell counts which is an indirect criterion. to assess susceptibility to mastitis. Data analyzed were weekly records (6448 somatic cell scores) out of 159 primiparous Holsteins and Holsteins x Normande cows raised at the INRA" Le Pin au Haras" experimental farm, France. Given the longitudinal structure of this data set, the analysis consists of modeling both the mean and the individual profiles. This was achieved via the use of mixed models including fixed effects for the average profiles and random effects for the adjusted individual profiles. As far as fixed effects are concerned, the main issue is to fit a time trend to the average profiles. For this, we employed the technique of fractional polynomials described in Royston and Altman (Appl. Stat. 43 (1994) 429–467) under several variance-covariance structures. The best second degree polynomial involved an intercept plus the time at the power (–1/3) (i.e., $t^{-1/3}$) plus the latter times the logarithm of the time (i.e., $t^{-1/3} \times \log(t)$). Regarding random effects, model comparisons involved random coefficient models, exponential stationary stochastic processes and heterogeneous variances. The models that simultaneously included all these three structures turned out to be the best. For instance, random coefficient models did not fit the variance function well, even when the degree of the polynomial was high. This phenomenon partly justified the introduction of heteroskedastic models.

longitudinal data / mixed models / fractional polynomials / robust estimators / somatic cell scores

Résumé – Analyse statistique des scores de cellules somatiques par la méthodologie des modèles mixtes appliquée à des données longitudinales. L'objectif de cette étude est d'analyser les comptages de cellules somatiques, qui constituent un critère indirect d'appréciation de la sensibilité aux mammites. Les données analysées étaient relatives à 6448 contrôles hebdomadaires de cellules somatiques du lait effectués sur 159 génisses de race Holstein et croisées Holstein x Normande, entretenues au domaine expérimental INRA-Le pin au Haras. Eu égard à la structure longitudinale

* Corresponding author: robert@germinal.toulouse.inra.fr

des données, l'analyse a consisté à modéliser les profils moyens et individuels de réponse. À ce propos, ont été mis en œuvre des modèles mixtes dont les effets fixes décrivent les profils moyens et les effets aléatoires les profils individuels. Concernant les effets fixes, on a utilisé la technique des polynômes fractionnaires, décrite par Royston et Altman [29] assortie de différentes structures de variance-covariance. Le meilleur ajustement est fourni par un polynôme de second degré comportant un terme constant, le temps à la puissance moins un tiers (i.e., $t^{-1/3}$) et ce même terme multiplié par le logarithme du temps (i.e., $t^{-1/3} \times \log(t)$). Concernant la partie aléatoire du modèle, les modèles mis en comparaison faisaient appel à une régression sur le temps à coefficients aléatoires, à des processus stochastiques stationnaires et à des variances hétérogènes. Ce sont les modèles qui incluaient simultanément ces trois structures qui se sont avérés les meilleurs. Ainsi, les modèles polynomiaux à degré élevé ne rendaient pas compte de l'évolution réelle de la variance avec le temps d'où le recours à des modèles hétéroscédastiques.

données longitudinales / modèles mixtes / polynômes fractionnaires / estimateurs robustes / scores de cellules somatiques

1. INTRODUCTION

Somatic cell count (SCC) has been widely advocated as an indicator trait for mastitis [17, 30]. In many countries, SCC is measured on a large scale in national milk recording systems (usually on a monthly basis) and used as an indirect criterion in genetic selection for mastitis resistance [17]. Genetic models for SCC are mostly based on lactation average somatic cell scores (SCS = log transformed SCC to achieve normality of distribution, [2]). This average of SCS at the individual level does not take into account the dynamics of somatic cell count in the course of lactation and also serial correlations among measurements made on the same individual.

The purpose of this study was to propose and compare models taking into account such phenomena. A class of models for such longitudinal data consists of mixed linear models including fixed effects for describing the average profiles and random effects for the adjusted individual profiles. This study was based on experimental data produced and collected under controlled environmental conditions and with accurate follow-up of SCS over time in each individual in order to fully explore the potential of these models and to make comparisons among them. In the materials and methods chapter, we will first describe the data set and secondly the statistical models with a separate presentation for means and variances. Mean models will be based on polynomial fitting on time with both conventional and fractional polynomials. Variance models aimed at describing the variance-covariance structure of SCS over time via random variables such as random regression coefficients and stochastic time processes, with or without heterogeneity of variances. The results are presented along the same pattern considering the means and variances separately. The paper finishes with a general discussion about the models considered here and possible alternatives.

2. MATERIALS AND METHODS

2.1. Data

The study was based on a survey conducted on an INRA experimental farm (Le Pin au Haras, Normandie, France) between 1998 and 1999 with the objective of assessing the relationship between somatic cell count and mastitis. Because the mean profiles of somatic cell scores between primiparous and multiparous were not the same [31], the data set used here was restricted to 159 primiparous purebred (Holstein cows) and F1 cows (Holstein x Normande crossbred), calving between 1998 and 1999. Available information consisted of 6448 SCS, i.e., one record per week and per cow over a period running from day 5 to day 305

Table I. Characteristics of the data set.

(a)	Number of animals	Number of animals	
	(number of observations)	(number of observations)	
	purebred	crossbred	
Calving year			
1998	36 (1438)	38 (1492)	74 (2930)
1999	19 (779)	66 (2739)	85 (3518)
	55 (2217)	104 (4231)	159 (6448)

(b)

	Number of animals	Number of animals	
	(number of observations)	(number of observations)	
	purebred	crossbred	
Calving season			
1: August–Sept.	26 (1055)	70 (2847)	96 (3902)
2: Oct.–Nov.	15 (622)	23 (958)	38 (1580)
3: Dec.–May	14 (540)	11 (426)	25 (966)
	55 (2217)	104 (4231)	159 (6448)

(c)

	Number of animals	Number of animals	
	(number of observations)	(number of observations)	
	purebred	crossbred	
Calving age			
(age at first calving)			
1: < 25 months	28 (1148)	24 (957)	52 (2105)
2: 25 to 31 months	14 (568)	27 (1111)	41 (1679)
3: > 31 months	13 (501)	53 (2163)	66 (2664)
	55 (2217)	104 (4231)	159 (6448)

after calving. In this data set, the animals can differ both in the number of records and in the length of the time intervals between them. About 40 records per animal were used for modeling the SCS curve all over the lactation period.

The distribution of the number of records and animals according to calving year (2 levels), calving season (3 levels) and calving age (3 levels) are shown for pure and crossbred heifers in Table I.

2.2. Statistical model

One of the most frequently used approaches in longitudinal data analysis is the linear mixed effects model [12] allowing for both a description of time trend and a specification of the correlation structure of the data.

Let y_{ij} be the jth measurement (j = 1, 2, ..., n_i) recorded on the ith animal (i = 1, 2, ..., I) at time t_{ij}. Models considered here are within the class of regression models, i.e.:

$$y_{ij} = x'_{ij}\beta + \varepsilon_{ij} \qquad (1)$$

where $\mathrm{x}'_{ij}\beta$ represents a linear combination of p explanatory variables (row vector x'_{ij} including discrete factors and/or continuous covariates) with unknown linear coefficients (vector β) and ε_{ij} is the random component. In matrix notation, this model can be written as follows:

$$y_i = \mathbf{X}_i\beta + \varepsilon_i \qquad (2)$$

where $\mathbf{y}_i = \{y_{ij}\}$, $\varepsilon_\mathrm{i} = \{\varepsilon_{ij}\}$, $\mathbf{X}_{i(n_i \times p)} = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, ...; \mathbf{x}_{in_i})$'.

We will assume that data are correlated ($Cov(y_{ij}, \mathrm{y}_{\mathrm{ij}'}) \neq 0$) and normally distributed so that $\varepsilon_i \sim N(\mathbf{O}, \mathbf{V}_i)$, the variance covariance matrix $\boldsymbol{V}_i = Var(\varepsilon_i)$ is not diagonal and is subject itself to some modeling.

Selecting a model is complex because the choice of the systematic part (fixed effects) depends on the variance-covariance structure of observations and vice versa. In practice, the strategy adopted is as follows: a structure of the variance-covariance matrix is assumed and a model is chosen after selection of some fixed effects; subsequently, assuming this model for the fixed part, different structures for the variance-covariance matrix are tested. In this paper, a slightly different approach is considered to reduce the dependency between choices at the two steps [24, 25]. This approach consists of making an inference on the fixed effects via robust estimators with respect to the structure imposed on the variance-covariance of observations. This robust approach is described in the paper of Liang and Zeger [13]. Then, after selection of the fixed effects, the second step consists of selecting and testing several variance-covariance structures of observations.

As far as fixed effects are concerned, the main issue is to adjust the data for the time trend. Conventional polynomials (with positive integer powers) are a classical choice for modeling the relationship between response variables and one or several continuous covariates. However, the curve does not usually fit the data well both at the low and high orders of these polynomials [16, 29]. 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 chose the technique of fractional polynomials, described in Royston and Altman [29]. This technique is just an extension of conventional polynomials but with real powers and is described in the "Models for the mean" section.

Different variance covariance structures (random coefficients, exponential stationary stochastic processes and heterogeneous variances) will be described and compared in the "Variance covariance structure" section.

Inference is based on the maximum likelihood (ML) and on residual likelihood procedures (REML, [18]) for the location and dispersion parameters respectively; and computations are made using the MIXED procedure of the SAS software [33].

2.3. Models for the mean

Description of the mean profile of somatic cell scores (SCS) requires to adjust the appropriate time trend for modeling the average trend of SCS during lactation and to select significant fixed effects (environmental factors, interactions between them and interactions between environmental factors and time trend). Time trend is modeled here via fractional polynomials, due to their simplicity, flexibility and parsimony. Fractional polynomials are an extension of conventional polynomials but with real powers.

2.3.1. Fractional polynomials

The family of fractional polynomials represents a linear combination of functions of time with real powers. Let t be a positive real covariable, $\mathbf{p} = (p_j; j = 0, 1, ..., m)$ a $(m + 1)$ vector of ordered powers called the vector of degree m and $\xi = (\xi_j; j = 0, 1, ..., m)$ the vector of the corresponding real coefficients. A fractional polynomial of degree m is defined as follows:

$$\phi_m(t, \xi, \mathbf{p}) = \sum_{j=0}^{m} \xi_j H_j(t) \qquad (3)$$

where $H_j(t) = t^{(p_j)}$ if $p_j \neq \mathrm{p}_{j-1}$ and $H_j(t) = \mathrm{H}_{j-1}(t)\ln(t)$ if $p_j = \mathrm{p}_{j-1}$. At the origin (for $j =$ = 0), $H_0(t) = 1$ and $p_0 = 0$.

In this last formula, $t^{(p_j)}$ represents the Box-Tidwell transformation [3] i.e., $t^{(p_j)} = t^{p_j}$ if $p_j \neq 0$ and $t^{(p_j)} = \ln t$ if $p_j = 0$.

For example, for $m = 3$ and $\mathbf{p} = (0, 0, 0, 4)$, the third order fractional polynomial is given by the function: $\phi_3(t;\ \xi,\ \mathbf{p}) = \xi_0 + \xi_1 \ln t + \xi_2 (\ln t)^2 + \xi_3 t^4$.

For modeling a data set using fractional polynomials, we need to determine the best value of m (degree of the polynomial) and of the power vector $\mathbf{p}$. The procedure for selecting the power ($\boldsymbol{p}$) and the degree of the fractional polynomial (m) is breifly described in Appendix 1.

2.3.2. Selection of other fixed effects

The model for other fixed effects (breed, calving year, calving season, calving age) 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 [13] and described by Robert-Granié et al. [24, 25]. This test is relatively insensitive to the structure of the variance covariance assumed for the data. The so-called "sandwich estimator" for $var(\hat{\beta})$ proposed by Liang and Zeger [13] is obtained by replacing $var(\mathbf{y_i})$ by $\mathbf{r_i}\mathbf{r_i'}$ where $\mathbf{r_i} = \mathbf{y_i} - \mathbf{X_i}\hat{\beta}$. The resulting estimator can then be shown to be consistent, as long as the mean is correctly specified in the model [36].

To that respect, the simplest choice consists of fitting β by ordinary least squares. Comparisons between robust and standard estimators are described in the paper of Robert-Granié et al. [25].

2.4. Variance covariance structure

Several ways are available for modeling individual profiles. Currently, the most common one relies on random coeffcient models [5, 15]. Random coefficient models are basically regression models incorporating random effects for the coefficients in order to explain the between subject component of variation observed in longitudinal data. Other tools may be envisioned e.g., stochastic time processes. In any case, one has to find a compromise between the quality of fit and the number of parameters used in the parametric functions for variances and covariances.

As presented by Diggle [4], ε_{ij} in model (1) can be decomposed as the sum of 3 sources of variation (between subjects, between times within a subject and measurement errors):

$$\varepsilon_{ij} = \sum_{k=1}^{K} z_{ijk}u_{ik} + w_i(t_{ij}) + e_{ij}\,.$$

The first term $(\sum_{k=1}^{K} z_{ijk}u_{ik})$ represents the additive effect of K random regression factors u_{ik} on covariable information z_{ijk} (usually a (k–1)th power of time) and which are specific to each ith individual. The second term $w_i(t_{ij})$ is a term sampled from copies of a stationary Gaussian process resulting in serial correlations between measurements of the same subject. The third term e_{ij} is a residual representing, either a pure measurement error for observations made at the same time or, as here, a pseudo measurement error estimated indirectly as a deviation from the parametric model. In matrix notation, this model can be written as follows:

$$\varepsilon_i = \mathbf{Z}_i\mathbf{u}_i + \mathbf{w}_i + \mathbf{e}_i$$

where $\mathbf{Z}_{i(ni \times K)} = (\mathbf{z}_{i1},\ \mathbf{z}_{i2},\ ...,\ \mathbf{z}_{in_i})'$, $\mathbf{z}_{ij(K \times 1)} = \{z_{ijk}\}$, $\mathbf{u}_{i(K \times 1)} = \{u_{ik}\}$ for $k = 1, 2, ..., K$, $\mathbf{w}_i = \{w_i(t_{ij})\}$ and $\mathbf{e}_i = \{e_{ij}\}$ for $j = 1, 2, ..., n_i$. We will assume that $\varepsilon_i \sim N(\mathbf{0}, \mathbf{V_i})$ with $\mathbf{V}_i = \mathbf{Z}_i\mathbf{G}\mathbf{Z_i'} + \mathbf{R_i}$ where $\mathbf{G}_{(\mathrm{K} \times \mathrm{K})} = Var(\mathbf{u}_i)$ is a symmetric positive definite matrix. For instance, for a linear regression $\mathbf{G} = \begin{pmatrix} g_{00} & g_{01} \\ g_{10} & g_{11} \end{pmatrix}$ where g_{00} refers to the variance of the intercept, g_{11} to the variance of the linear regression coefficient and g_{01} to their covariance. $\mathbf{R_i}$ has the following structure in the general case: $\mathbf{R}_i = \sigma^2\mathbf{H}_i + \sigma_e^2\ \mathbf{I_{n_i}}$, where $\sigma_e^2\,\mathbf{I_{n_i}} = var(\mathbf{e}_i)$, and for stationary Gaussian simple processes, σ^2 is the variance of each $w_i(t_{ij})$ and $\mathbf{H}_i = \{h_{ij,\ ij'}\}$ the $(n_i \times n_i)$ correlation matrix among them such that $h_{ij,\ ij'} = f(\rho, \mathrm{d}_{ij,ij'})$ can be written as a function f of a real positive number ρ and of the absolute time separation $d_{ij,ij'} = |t_{ij} - t_{ij'}|$ between measurements j and j' made on the individual i. A classical example of such functions is the power function: $f(\rho, d) = \rho^d$. Notice that for equidistant intervals, this power function is equivalent and reduces the autoregressive process to a first order.

We can extend this model to take into account heterogeneous variances both at the temporal (in **G** matrix) and residual (in **R** matrix) levels [6, 8, 10, 20, 21, 23, 32] which enlarges the range of potentially useful models for longitudinal data analysis.

A convenient and parsimonious procedure to handle heterogeneous variances is to model them linearly via a log link function [6, 7, 32]. Here, this model for log-residual variances can be simply written as: $\ln \sigma^2_{e_{ij}} = a_0 + a_1 t_{ij}$ where a_0 and a_1 are unknown real coefficients and t_{ij} represents the stage of lactation (in days) of the jth measurement recorded on the ith animal.

The aim of this part was to compare these different approaches (conventional polynomials, fractional polynomials, stochastic processes, heterogeneous variances) and to evaluate their behavior (separately or by combining them) for modeling individual profiles. The models compared are presented in Table II. We shall also rely on graphical diagnosis tools to assess how models based on variances over time agree with empirical values.

The selection of random effects was based on the likelihood ratio tests (REML version) when the models were nested and based on the Akaike criteria (AIC) for the other models. For nested models, the null hypothesis (H_0) can be described as a point hypothesis with parameter values on the boundary of the parameter space which implies some change in the asymptotic distribution of the likelihood ratio statistic under H_0 [34, 35]. Actually, the asymptotic distribution of the likelihood ratio test statistitic (REML version) under the null hypothesis is a mixture $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ of the usual chi-square with one degre of freedom χ_1^2 and of a Dirac (probability mass of one) at zero (usually noted χ_0^2) with equal weights. This results in a P-value which is half the standard one i.e., P-value $= \frac{1}{2}Pr[\chi_1^2 > \delta]$ where δ represents a restricted likelihood ratio statistic of two tested models (see [35] for a similar application). For non nested models, the best model will be the one having the highest value of the Akaike criteria (or the lowest value of –2 AIC).

3. RESULTS

3.1. Models for the mean

3.1.1. Plot of data

With longitudinal data, an obvious first graph to consider is the empirical mean profile of somatic cell scores according to the stage of lactation (in days). This graph (Fig. 1) illustrates a sharp decrease in the first thirty days and, thereafter, a gradual increase until the end of lactation. Variation in the shape and level of the SCS pattern is related to udder infection status and to individual cows [31].

3.1.2. Time trend

The best fractional polynomial model was found with the model defined in (2) and a variance-covariance structure $\mathbf{V_i} = \mathbf{I_{n_i}}\,\sigma_e^2$. When assessing competing fractional polynomial models of degree 1 or 2, it is often informative to plot the gain G (see Appendix I) against the chosen powers

Table II. Models with different variance covariance structures.

Groups	Models	$\mathbf{Z_i}$	$\mathbf{G}$	$\mathbf{R_i}$	# par
A	[0]	$\mathbf{O}_{n_i}$		$\sigma_e^2\mathbf{I}_{n_i}$	1
	[1]	$1_{\mathbf{n_i}}$	g_{00}	$\sigma_e^2\mathbf{I}_{n_i}$	2
	[2]	$(1_{\mathbf{n_i}}, \mathbf{t}_i)$	$g_{00}\, g_{01}$ $g_{01}\, g_{11}$	$\sigma_e^2\mathbf{I}_{n_i}$	4
	[3]	$(1_{\mathbf{n_i}}, \mathbf{t}_i, \mathbf{t}_i^2)$	$g_{00}\, g_{01}\, g_{02}$ $g_{01}\, g_{11}\, g_{12}$ $g_{02}\, g_{12}\, g_{22}$	$\sigma_e^2\mathbf{I}_{n_i}$	7
B	[4]	$\mathbf{O}_{\mathbf{n_i}}$		$\sigma^2\mathbf{H}_i$	2
	[5]	$\mathbf{O}_{\mathbf{n_i}}$		$\sigma^2\mathbf{H}_i + \sigma_e^2\mathbf{I}_{n_i}$	3
C	[6]	$1_{\mathbf{n_i}}$	g_{00}	$\sigma^2\mathbf{H}_i + \sigma_e^2\mathbf{I}_{n_i}$	4
	[7]	$(1_{\mathbf{n_i}}, \mathbf{t}_i)$	$g_{00}\, g_{01}$ $g_{01}\, g_{11}$	$\sigma^2\mathbf{H}_i + \sigma_e^2\mathbf{I}_{n_i}$	6
	[8]	$(1_{\mathbf{n_i}}, \mathbf{t}_i, \mathbf{t}_i^2)$	$g_{00}\, g_{01}\, g_{02}$ $g_{01}\, g_{11}\, g_{12}$ $g_{02}\, g_{12}\, g_{22}$	$\sigma^2\mathbf{H}_i + \sigma_e^2\mathbf{I}_{n_i}$	9
D	[9]	$1_{\mathbf{n_i}}$	g_{00}	$\ln \sigma_{eij}^2 = a_0 + a_1\, t_{ij}$	3
	[10]	$(1_{\mathbf{n_i}}, \mathbf{t}_i)$	$g_{00}\, g_{01}$ $g_{01}\, g_{11}$	$\ln \sigma_{eij}^2 = a_0 + a_1\, t_{ij}$	5
	[11]	$(1_{\mathbf{n_i}}, \mathbf{t}_i, \mathbf{t}_i^2)$	$g_{00}\, g_{01}\, g_{02}$ $g_{01}\, g_{11}\, g_{12}$ $g_{02}\, g_{12}\, g_{22}$	$\ln \sigma_{eij}^2 = a_0 + a_1\, t_{ij}$	8
E	[12]	$1_{\mathbf{n_i}}$	g_{00}	$\sigma^2\mathbf{H}_i + \sigma_{eij}^2$ with $\ln \sigma_{eij}^2 = a_0 + a_1\, t_{ij}$	5
	[13]	$(1_{\mathbf{n_i}}, \mathbf{t}_i)$	$g_{00}\, g_{01}$ $g_{01}\, g_{11}$	$\sigma^2\mathbf{H}_i + \sigma_{eij}^2$ with $\ln \sigma_{eij}^2 = a_0 + a_1\, t_{ij}$	7
	[14]	$(1_{\mathbf{n_i}}, \mathbf{t}_i, \mathbf{t}_i^2)$	$g_{00}\, g_{01}\, g_{02}$ $g_{01}\, g_{11}\, g_{12}$ $g_{02}\, g_{12}\, g_{22}$	$\sigma^2\mathbf{H}_i + \sigma_{eij}^2$ with $\ln \sigma_{eij}^2 = a_0 + a_1\, t_{ij}$	10

# par: the number of estimated dispersion parameters.
A: Random coefficient models; B: Time process model (+ possibly measurement error); C: Random coefficient models + time process + measurement error; D: Random coefficient models + heteroskedastic measurement error; E: Random coefficient models + time process + heteroskedastic measurement error.
t_i: is the $n_i \times 1$ vector of lactation stage at which measurements are made on individual i, $H_i = \{h_{i,tt'} = \rho^{|t_i - t_i'|}.\}$

(Fig. 2); for $m = 1$, here denoted p_1 rather han p, so $\mathbf{p} = (p_0 = 0, p_1)$. For the $m = 2$ models, we have $\mathbf{p} = (p_0 = 0, p_1, p_2)$, the gain G may be plotted against p_1 on the same graph as a sheaf of curves indexed by the chosen values p_2. For $m = 1$, the plotted curve (p_2 =. in Fig. 2) essentially depicts the profile deviance function for p_1 when p_1 is

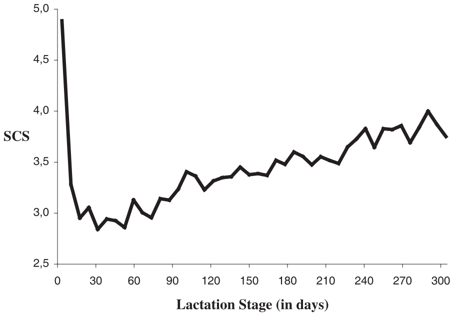

Figure 1. Mean profile of somatic cell score during lactation.

regarded as a parameter of the non-linear model: $\xi_0 + \xi_1 t^{p_1}$. If a curve has a peak, the value p_1 corresponding to the maximum G is the MLE $\hat{p}_1$, so the plot gives an idea of how close $\tilde{p}_1$ is to $\hat{p}_1$ (see Appendix I for the definition of $\tilde{p}_1$). For m = 2, the family of curves indexed by p_2 illustrates graphically the power vectors (0, p_1, p_2) which give high values of G (indicating a good fit) and those which are associated with a less good fit.

The results for the first two degrees (m = 1, 2) are shown in Figure 2. Gain is plotted against p_1 values with different curves for p_2 values. The best p and sub-optimal p value models are shown in Table III. 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 G of 102.28. 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 (see Tab. III). The p_1 and p_2 values

Table III. Gain values for the best and sub-optimal fractional polynomial models: m = 2 and power vector $\mathbf{p}$ = (0, p_1, p_2).

p_1	p_2	Deviance	Gain
–1/3	–1/3	23848.95	102.28
–1/2	0	23849.42	101.80
–1/2	–1/3	23851.00	100.22
–1	1/3	23851.17	100.05
–1/3	0	23851.63	99.59
–1	1/2	23851.93	99.29
–1/2	–1/2	23855.38	95.85
–1 0	0	23855.83	95.40
–1/2	1/3	23855.87	95.35
–1/2	1/2	23861.16	90.07

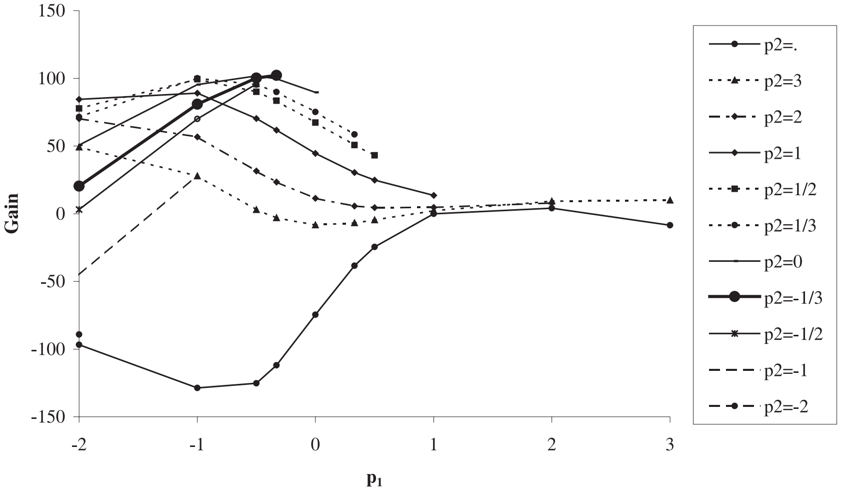

Figure 2. Gain plotted against p_1 for the fractional models: $\phi_1(t; \xi=(\xi_0, \xi_1); \mathrm{p} = (0, p_1))$ and for $\phi_2(\mathrm{t}; \xi = (\xi_0, \xi_1, \xi_2); \mathrm{p} = (0, p_1, p_2))$ with different values of p_2.

found here were in good agreement with a non linear fit of the data for which $\hat{p}_1 = \hat{p}_2 = -0.30$ with a gain of 102.42. This was achieved by using the procedure nlin of the SAS software [33].

Fractional polynomial models of degree 3 were tested but the gain obtained with the best model with $m = 3$ ($G = 103.2$) was not significantly different from that obtained with the best model with $m = 2$ ($G = 102.28$).

Therefore, we recommend using the second order fractional polynomials, found with this data set: $\phi_2(t; \xi, \mathbf{p}) = \xi_0 + \xi_1 t^{-1/3} + \xi_2 t^{-1/3} \ln(t)$ with $\mathbf{p} = (0, -1/3, -1/3)$.

Incidentally, a quartic conventional polynomial (for which $G = 50.9$) with $m = 4$ and $\mathbf{p} = (0, 1, 2, 3, 4)$ was unable to produce a fit as good as that of the best second order fractional polynomial model.

Figure 3 shows the mean profile of somatic cell scores fitting by different functions: (1) the best fractional polynomial found previously with $m = 2$ and $\mathbf{p} = (0, -1/3, -1/3)$; (2) the best conventional polynomial (with integer powers) found among all conventional polynomials ($m = 3$ and $\mathbf{p} = (0, 1, 2, 3)$); (3) the Ali and Schaeffer function [1] defined by: $f_{(t)}=a_0 + a_1\frac{t}{305} + a_2\left(\frac{t}{305}\right)^2 + a_3 \ln\left(\frac{305}{t}\right) + a_4\left(\ln\left(\frac{305}{t}\right)\right)^2$ and often used to fit the mean profile of the milk production in dairy cattle. This function can also be interpreted as a fractional polynomial of degree $m = 4$ with $\mathbf{p} = (0, 0, 0, 1, 2)$. The last model leads to a deviance of $D = 23847.3$ and a gain of $G = 103.92$ sightly better (but not significantly) than the one obtained with the previously selected best second order fractional polynomial. Moreover, the second order fractional polynomial involves less parameters and fits seemingly better the right part of the curve at the end of lactation (see Fig. 3). The best conventional polynomial (degree 3 and integer powers) found among all conventional polynomials, shows the usual waviness and end-effects that are often associated with high degree polynomials. The conventional polynomial fits data very poorly at the beginning of lactation. Finally, the fractional polynomial described somatic cell scores almost as well as the Ali & Schaeffer specification.

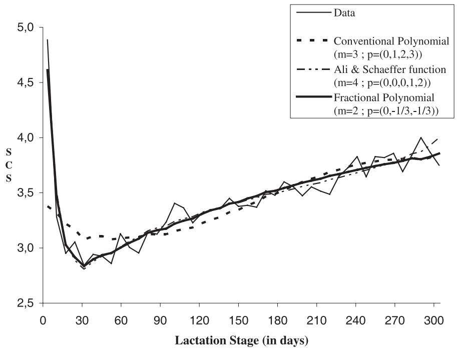

Figure 3. Mean profile of somatic cell scores fitted by different functions.

3.1.3. *Selection of fixed effects using robust estimators*

As explained in the "Models for the mean" section, the fixed effects were selected using robust estimators [13]. Practically, the resulting standard errors can be requested in the SAS Mixed procedure by adding the option "empirical" in the proc mixed statement. Table IV presents the value of the F-type statistic described in Littell et al. [14] on page 502 and the P-value associated with each fixed factor considered. In view of these results, the list of fixed effects retained in the model was: a second order fractional polynomial in time, breed, calving year, calving season, calving season × breed and calving season × year interactions. The interactions between factors and the time function were not significant.

3.2. Variance covariance structure

For a given mean model, there are different competing variance covariance structures

Table IV. Selection of fixed effects by robust estimators.

Fixed effects	Value of F-test	*P*-value
$t^{-1/3}$	25.73	< 0.0001*
$\ln(t) \times t^{-1/3}$	64.97	< 0.0001*
Breed (2 levels)	2.43	0.1215
Calving year (2 levels)	0.35	0.5538
Calving season (3 levels)	1.38	0.2539
Calving age (3 levels)	0.32	0.7291
Breed × Calving year	1.97	0.1631
Breed × Calving season	3.54	0.0316*
Breed × Calving age	0.65	0.5228
Calving year × Calving season	3.82	0.0242*
Calving year × Calving age	0.92	0.4027
Calving season × Calving age	0.54	0.7044
$t^{-1/3}$ × Breed	0.12	0.7238
$\ln(t) \times t^{-1/3}$ × Breed	1.95	0.1622
$t^{-1/3}$ × Calving year	0.36	0.5461
$\ln(t) \times t^{-1/3}$ × Calving year	0.43	0.5124
$t^{-1/3}$ × Calving season	1.44	0.2362
$\ln(t) \times t^{-1/3}$ × Calving season	0.66	0.5169
$t^{-1/3}$ × Calving age	0.16	0.8487
$\ln(t) \times t^{-1/3}$ × Calving age	0.31	0.7339

*: significant at the level $\alpha = 5\%$.

that can be compared. The models considered in this application are presented in Table V. These models do not constitute an exhaustive list, since there are other possible specifications.

The results in Table V show large values for the likelihood ratio statistics. The models included in each group (A, B, C, D and E) are nested. We can then compare these models within each group using the likelihood ratio test and between groups, the Akaike criterion can be used to compare the best models from each group.

The fractional polynomial with $m = 2$ and $\mathbf{p} = (0, -1/3, -1/3)$ were also considered for fitting the individual part of the model. This model, similar to model [3] of Table V, including 7 parameters, had a value of –2 log-likelihood equal to 17508.00 and of –2 AIC equal to 17522.00. A conventional second degree random coefficient model (model [3] of Tab. V) appears to be better than the fractional polynomial. This example clearly shows that functions selected at the expectation level are not necessarily adequate for the covariance level.

The model finally accepted is model [1]: a random coefficient model (with a second degree polynomial function in time to describe the individual part) plus a stationary gaussian simple process (with a power function) and a heteroskedastic measurement error (the logarithm of the residual variance is modeled by a linear function of time).

In addition, a graphical diagnosis was performed to check whether the 3 best models generate a variance function against time (in weeks) which is close enough to the empirical one. Figure 4 shows variance functions obtained with different models considered in Table V. The dotted curve ("Empirical variance plotted against time" in Fig. 4) represents the variance function obtained from a fixed model (model [0]) where residuals are assumed independant (for each week, the variance of residuals has been computed). The bold curve ("Smoothed empirical variance" in Fig. 4) is obtained by smoothing the observed squared residuals against time; and the last three curves represent the variance function of models [12], [13] and [14], described in Tables II and V. On the basis of this graph, model [12] turns out to fit the smoothed empirical variance better. The variance function generated by model [13] shows the usual pattern of the linear random regression models whereas the variance function under model [14] suffers from a waviness pattern (usually observed with the adjustment of conventional polynomials with integer powers and m ≥ 2).

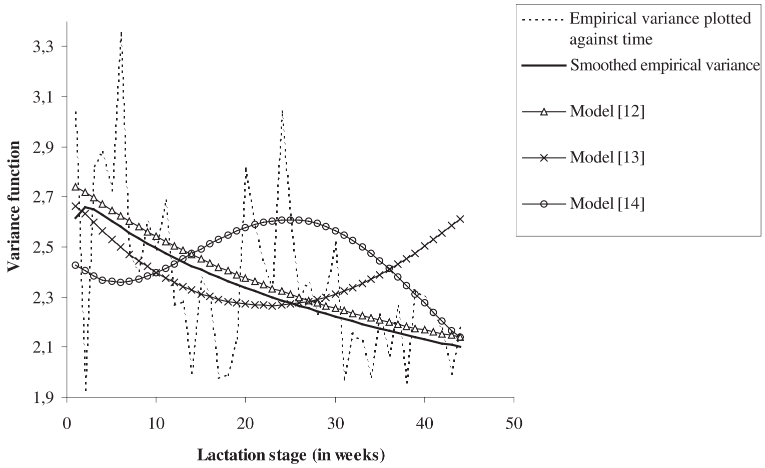

Figure 4. Variance functions.

Table V. Model selection statistics for alternative variance-covariance structures.

Groups	Models[a]	# par	–2RL[b]	Comparisons	Δ[–2RL][c]	Distr[d]	*P*-value	–2AIC[e]
A	[0]	1	23884.2					23886.2
	[1]	2	18669.9	[1]–[0]	5214.3	0:1	< 0.0001	18673.9
	[2]	4	17742.7	[2]–[1]	927.2	1:2	< 0.0001	17750.7
	[3]	7	17318.0	[3]–[2]	424.7	2:3	< 0.0001	17332.0
B	[4]	2	17536.0					17540.0
	[5]	3	16639.4	[5]–[4]	896.6	0:1	< 0.0001	16645.4
C	[6]	4	16595.4					16603.4
	[7]	6	16568.7	[7]–[6]	26.7	1:2	< 0.0001	16580.7
	[8]	9	16537.2	[8]–[7]	31.5	2:3	< 0.0001	16555.2
D	[9]	3	18320.5					18326.5
	[10]	5	17401.3	[10]–[9]	919.2	1:2	< 0.0001	17411.3
	[11]	8	16960.9	[11]–[10]	440.4	2:3	< 0.0001	16976.9
E	[12]	5	16337.8					16347.8
	[13]	7	16321.5	[13]–[12]	16.3	1:2	< 0.0001	16335.5
	[14]	10	16297.7	[14]–[13]	23.8	2:3	< 0.0001	16317.7

A: Random coefficient models; B: Time process model (+ possibly measurement error); C: Random coefficient models + time process + measurement error; D: Random coefficient models + heteroskedastic measurement error; E: Random coefficient models + time process + heteroskedastic measurement error.
[a]: All these models are described in Table II; [b]: –2RL = –2 × Log restricted likelihood; [c]: Likelihood ratio statistics of two tested models; [d]: Asymptotic distribution of the likelihood ratio under the null hypothesis: mixture of chi-squares (for instance 1:2 represents a mixture in equal proportions of two chi-squares with 1 and 2 degrees of freedom respectively); [e]: –2AIC = –2 × the Akaike criterion.

4. DISCUSSION AND CONCLUSION

Relative to current genetic models based on lactation average of SCS, the models for test day observations should account better for short term environmental variation and allow using all information without restriction on the number of records available or length of time intervals. To that respect, test day models may also account more precisely for short time variation of SCC than average lactation models and be more efficient in predicting clinical cases and infections in general [31].

The use of models including continuous covariates is widespread but it has long been recognized that conventional polynomials often fit data poorly [16, 29]. However, it seems that few low dimensional parametric alternatives to, or extensions of, conventional polynomials have been suggested. Existing alternatives such as cubic splines, and non parametric smoothers often work well but also have drawbacks: they are computationally intensive, the application of theory of these methods is often difficult (choice of the number of knots, choice of a parameter to control the degree of smoothing), they do not yield compact expressions for prediction and the coefficients do not lend themselves to mechanistic interpretation. Eventually, the non parametric methods as some conventional polynomials, generate some artificial waviness because these methods tend to stick to the data.

As far as models for means are 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 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.

Regarding the random part of the model, we showed that functions selected at the expectation level are not necessarily adequate for the covariance level; this contradicts standard specifications used in most genetic test-day models. In many studies especially in animal breeding, the authors assume the same regression structure on the fixed and random effects. This is neither mandatory in theory nor desirable in practice, since variation between populations and between subjects within populations do not necessarily have the same pattern. In practice, the order of polynomials for fitting the random part of the model (adjusted profiles) is usually lower than that of the fixed part (population trend). Petim-Batista et al. [19] have compared several polynomial adjustments on each part of the genetic model of somatic cell scores (fixed effects, genetic and permanent environmental parts). They eventually retained a model with the Ali and Schaeffer function for the fixed effects and the permanent environmental effects; and a fractional polynomial with $m = 2$ and $\mathbf{p} = (0, -1/3, -1/3)$ for the genetic effects. This application clearly illustrates again that the function selected for the mean profile is not necessarily the same as the one selected for the individual profiles. Similar results are found in Robert-Granié et al. [26].

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 model is not an easy one since there is a strong dependency between the choices of the mean and the covariance structures. Eventually, this choice should be based not only on the usual information criteria (AIC, BIC, DIC) but also in relation to the final objective for adjusting SCS (indicator of mastitis) and their genetic variation and interpretation. In this application, we show that the model selected with the graphic tool of diagnosis (variance functions, Fig. 4) was not the same as that obtained when using the likelihood ratio tests (comparison models, Tab. V). Further analyses are needed to compare models not only on a criterion based on variance function but also by analyzing and comparing the empirical correlation function with those of several models considered. Furthermore, more analysis is needed to extend model comparisons by considering models with stochastic processes of higher order and by considering complete heteroskedastic models (for instance, heterogeneous individual and residual variances, [9, 22, 25, 26]).

Further analysis is needed to include genetic and permanent environmental effects into the model and to predict mastitis occurence based on SCS profiles. For this last point, new promising approaches have been developped, e.g. an extension of the multiprocess Kalman Filter [11]. This technique can be relevant for SCS and allows to provide probabilities of mastitis and hopefully will be able to detect mastitis earlier. An alternative would also be to consider pure non linear mixed models based on different typical functions that describe SCS subject patterns (see e.g. [27, 28]). This analysis is in fact the first stage of the study. The second stage would be to use the individual profiles of SCS to predict the occurence of mastitis.

ACKNOWLEDGEMENTS

The authors wish to thank H. Larroque (INRA) and J.J. Colleau (INRA) for providing the data set.

REFERENCES

[1] Ali T.E., Schaeffer L.R., Accounting for covariances among test day milk yields in dairy cows, Can. J. Anim. Sci. 67 (1987) 637–644.

[2] Ali A.K.A., Shook G.E., An optimum transformation for somatic cell concentration in milk, J. Dairy Sci. 63 (1980) 487–490.

[3] Box G.E.P., Tidwell P.W., Transformation of the independent variables, Technometrics 4 (1962) 531–550.

[4] Diggle P.J., An approach to the analysis of repeated measurements, Biometrics 44 (1988) 959–971.

[5] Diggle P.J., Liang K.Y., Zeger S.L., Analysis of longitudinal data, Oxford Science Publications, Clarendon Press, Oxford, 1994.

[6] Foulley J.L., Gianola D., San Cristobal M., Im S., A method for assessing extent and sources of heterogeneity of residual variances in mixed linear models, J. Dairy Sci. 73 (1990) 1612–1624.

[7] Foulley J.L., San Cristobal M., Gianola D., Im S., Marginal likelihood and Bayesian approaches to the analysis of heterogeneous residual variances in mixed linear Gaussian models, Comput. Stat. Data Anal. 13 (1992) 291–305.

[8] Foulley J.L., Quaas R.L., Thaon d'Arnoldi C., A Link function approach to heterogeneous variance components, Genet. Sel. Evol. 30 (1998) 27–43.

[9] Foulley J.L., Robert-Granié C., Heteroskedastic random coefficient models, 7th World Congress on Genetics Applied to Livestock Production, Montpellier, France, August 19-23, 32 (2002), pp. 157–160.

[10] Jaffrezic F., White I.M.S., Thompson R., Hill W.G., A link function approach to model heterogeneity of residual variances over time in lactation curve analysis, J. Dairy Sci. 83 (2000) 1089–1093.

[11] Korsgaard I.R., Lovendahl P., An introduction to multiprocess class II mixture models, 7th World Congress on Genetics Applied to Livestock Production, Montpellier, France, August 19-23, 32 (2002), pp. 185–188.

[12] Laird N.M., Ware J.H., Random effects models for longitudinal data, Biometrics 38 (1982) 963–974.

[13] Liang K.Y., Zeger S.L., Longitudinal data analysis using generalized linear models, Biometrika 73 (1986) 13–22.

[14] Littell R.C., Milliken G.A., Stroup W.W., Wolnger R.D., SAS System for Mixed Models, SAS institute Inc., 1996.

[15] Longford N.T., Random coefficients models, Clarendon Press, Oxford, 1993.

[16] McCullagh P., Nelder J.A., Generalized linear models, Chapman and Hall, London, 1989.

[17] Mrode R.A., Swanson G.J.T., Genetic and statistical properties of somatic cell count and its suitability as an indirect means of reducing the incidence of mastitis in dairy cattle, Anim. Breed. Abs. 64 (1996) 847–857.

[18] Patterson H.D., Thompson R., Recovery of interblock information when block sizes are unequal, Biometrika 58 (1971) 545–554.

[19] Petim-Batista F., Foulley J.L., Robert-Granié C., Silvestre A., Colaço J., SCS analysis in Portuguese dairy cows using random coefficients models, 7th World Congress on Genetics Applied to Livestock Production, Montpellier, France, August 19-23, 32 (2002), pp. 227–230.

[20] Robert C., Foulley J.L., Ducrocq V., Genetic variation of traits measured in several environments. I. Estimation and testing of homogeneous genetic and intraclass correlations between environments, Genet. Sel. Evol. 27 (1995) 111–123.

[21] Robert C., Foulley J.L., Ducrocq V., Genetic variation of traits measured in several environments. II. Inference on between-environment homogeneity of intraclass correlations, Genet. Sel. Evol. 27 (1995) 125–134.

[22] Robert-Granié C., Ducrocq V., Foulley J.L., Heterogeneity of variance for type traits in the Montbeliarde cattle breed, Genet. Sel. Evol. 29 (1997) 545–570.

[23] Robert-Granié C., Bonati B., Boichard D., Barbat A., Accounting for variance heterogeneity in French dairy cattle genetic evaluation, Livest. Prod. Sci. 60 (1999) 343–357.

[24] Robert-Granié C., Foulley J.L., Inférence robuste sur les effets fixes en modèle linéaire mixte pour l'analyse de données répétées, XXXIII^{es} Journées de Statistiques, Nantes, France, 2001.

[25] Robert-Granié C., Heude B., Foulley J.L., Modelling the growth curve of Maine-Anjou beef cattle using heteroskedastic random coefficients models, Genet. Sel. Evol. 34 (2002) 423–445.

[26] Robert-Granié C., Maza E., Rupp R., Foulley J.L., Use of fractional polynomial for modelling somatic cell scores in dairy cattle, 7th World Congress on Genetics Applied to Livestock Production, Montpellier, France, August 19-23, 32 (2002), pp. 153–156.

[27] Rodriguez-Zas S.L., Gianola D., Shook G.E., Evaluation of models for somatic cell score lactation patterns in Holsteins, Livest. Prod. Sci. 67 (2000) 19–30.

[28] Rodriguez-Zas S.L., Gianola D., Shook G.E., An approximate Bayesian analysis of somatic cell score curves in Holsteins, Acta Agric. Scand. 50 (2000) 291–299.

[29] Royston P., Altman D.G., Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling, Appl. Stat. 43 (1994) 429-467.

[30] Rupp R., Boichard D., Genetic parameters for clinical mastitis, somatic cell score, production, udder type traits, and milking ease in first lactation Holsteins, J. Dairy Sci. 82 (1999) 2198–2204.

[31] Rupp R., Analyse Génétique de la résistance aux mammites chez les ruminants laitiers, Ph.D. thesis, INA-PG/INRA, 2000.

[32] San Cristobal M., Foulley J.L., Manfredi E., Inference about multiplicative heteroskedastic components of variance in a mixed linear Gaussian model with an application to beef cattle breeding, Genet. Sel. Evol. 25 (1993) 3–30.

[33] SAS® Institute Inc., Cary NC: SAS® institute Inc., SAS/STAT Software, version 8, 1999.

[34] Self S.G., Liang K.Y., Asymptotic properties of maximum likelihood estimation and likelihood ratio tests under nonstandard conditions, J. Am. Stat. Assoc. 82 (1987) 605–610.

[35] Stram D.O., Lee J.W., Variance components testing in the longitudinal mixed effects models, Biometrics 50 (1994) 1171–1177.

[36] Verbeke G., Molenberghs G., Linear mixed models for longitudinal data, Springer Verlag, New-York, 2000.

APPENDIX I: BASIC THEORY OF FRACTIONAL POLYNOMIALS

Conditional on given values of m (degree of the fractional polynomial) and $\mathbf{p}$ (power vector), $\phi_m(t;\, \xi,\, \mathbf{p})$ has the form of a linear predictor in terms of the covariate $H_j(t)$ and of the parameter ξ_j. Viewed thus, $\phi_m(t;\, \xi,\, \mathbf{p})$ is a particular suitable candidate for modeling the time trend, the statistical properties of linear models being of course better (or easier to establish) than those of non linear models. It is worth considering the families $\phi_1(.)$ an $\phi_2(.)$ specifically, Royston and Altman [29] have found that models with degrees higher than 2 are rarely required in practice. Fractional polynomials with $m \leq 2$ offer many potential improvements in fit compared with conventional polynomials (polynomials with integer powers); several examples are presented in the paper of Royston and Altman [29]. So, for modeling a data set using fractional polynomials, we need to determine the best value of m and the power vector $\mathbf{p}$.

Suppose that the elements of $\mathbf{p}$ are allowed to vary continuously (rather than being restricted to a fixed set), then $\phi_m(t, \xi, \mathbf{p})$ is a non-linear model with parameters $(\xi, \mathbf{p})$. Then, the quantity $D(m, \xi, \mathbf{p}) - D(m, \xi, \hat{\mathbf{p}})$ where $D = -2 \times$ log-likelihood and $\hat{\mathbf{p}}$, the maximum likelihood estimate (MLE) of $\mathbf{p}$, has an asymptotic chi-square distribution with m ($2m + 1$ minus $m + 1$) degrees of freedom. In practice, Royston and Altman [29] have shown that $\mathbf{p}$ can be restricted to m values selected from a fixed set P of powers (usually fractions but not necessarily). Here, we took $P = \{-2, -1, -1/2, -1/3, 0, 1/3, 1/2, ..., \max(3, m)\}$. And this set is sufficiently rich to cover many practical cases adequately. Let $\tilde{\mathbf{p}}$ the power vector associated with the model of lowest deviance over the restricted parameter space based on P, its deviance $D(m, \xi, \tilde{\mathbf{p}})$ is larger than $D(m, \xi, \hat{\mathbf{p}})$, so that $D(m, \xi, \mathbf{p}) - D(m, \xi, \tilde{\mathbf{p}})$ can be viewed as a conservative test for a given value of $\mathbf{p}$. That is why Royston and Altman [29] proposed to select models for which the difference is lower or equal to the 90% quantile of a chi-square distribution with m degrees of freedom.

Specifically, when $m = 1$, the criterion $D(1, \xi, 1) - D(1, \xi, \tilde{\mathbf{p}}) > \chi^2_{1;\,0.90}$ represents a test with a significance level of about 10% for $p = 1$ (linearity) against $p \neq 1$ (monotonic alternatives) which may be used in an initial investigation of non linearity.

In practice, for general m, we suggest choosing models with values of $\mathbf{p}$ such that $D(m, \xi, \mathbf{p}) - D(m, \xi, \tilde{\mathbf{p}}) < \chi^2_{1;\,0.90}$ as the best fitting among those of degree m.

When deciding whether models with degree m are adequate or whether degree $m + 1$ is required, two extra parameters (a power and a regression coefficient) are estimated when m is increased by 1. Therefore, $D(m, \xi, \hat{\mathbf{p}}) - D(m + 1, \xi, \hat{\mathbf{p}})$ is asymptotically distributed as a chi-square with 2 degrees of freedom when the degree m model is adequate ($\hat{\mathbf{p}}$ refers implicitly to degree m or to degree $m + 1$ as appropriate).

So, the criterion $D(m, \xi, \tilde{\mathbf{p}}) - D(m + 1, \xi, \tilde{\mathbf{p}}) < \chi^2_{2;\,0.90}$ (= 4.7) is used for preferring models with degree $m + 1$ to those with degree m.

Usually the results are presented as a "gain" G which corresponds to the decrease in deviance from a straight line model: $G = G(m, \xi, \mathbf{p}) = D(1, \xi, 1) - D(m, \xi, \mathbf{p})$.

Since the gain G moves in the opposite direction to the deviance D, a larger gain indicates a better fit.

Once m and acceptable models of degree m have been selected, the final choice must depend mainly on the appearance of the curves in relation to the data, especially at the extreme values of the covariate (t).

To access this journal online:
www.edpsciences.org