Modelling Lactation Curves with Cubic Splines

I.M.S. White and S. Brotherstone
Institute of Cell, Animal and Population Biology
University of Edinburgh, West Mains Road, Edinburgh EH9 3JT

Abstract

Various approaches have been used to model lactation curves, mostly involving the fitting of parametric curves with fixed coefficients. Recently these models have been extended by allowing the coefficients to be random, but the resulting models still require the specification of an underlying parametric curve.

The fitting of splines represents a fully non-parametric approach to the problem. Cubic smoothing splines have recently been used to model growth curves, and they can easily be incorporated into a mixed-model framework. The potential for the use of splines in modelling lactation curves is explored with a simple example.

Introduction

Dairy cattle test day data provide an example of longitudinal data, or repeated measures, the essential feature of which is the presence of correlations between measurements (tests) on the same animal. Both genetic and environmental covariances need to be taken into account. Various methods have been proposed for analysing such data, ranging from relatively simple curve fitting through to a full multivariate analysis. In the genetic context, the latter amounts to treating the measurements at successive times as separate but correlated traits. All longitudinal methods of analysis can be regarded as derived from a model in which the data vector, arranged by trait within animal, has covariance matrix $I \otimes \Sigma$, or $A \otimes \Sigma$ for an animal model, with the different methods making more or less stringent assumptions about Σ.

Two curve-fitting methods, random regressions and spline fitting, use patterned covariance matrices in the analysis of longitudinal data, in the first case quite explicitly, in the second case less so. Spline curves have recently been recommended for statistical modelling by Verbyla (1997), on the basis that they are flexible and straightforward to fit. This paper describes our experience in fitting splines to test day records. First we outline the basic mathematical theory of splines and show how they can be fitted by methods (BLUP and REML) familiar to animal breeders.

Cubic splines

Natural cubic splines (NCS) are used for interpolation and nonparametric regression, and provide a more flexible class of curves than polynomials. A NCS with knots (i.e. data points) $t_1 < t_2 < ... < t_n$ can be represented as

$$\Phi_0 + \Phi_1 t + \tfrac{1}{6} \sum_{j=1}^{n} c_j (t - t_j)_+^3$$

where $\sum c_j = \sum c_j t_j = 0$, and $y_+^3 = y^3$ if $y > 0$, $y_+^3 = 0$ if $y \leq 0$. From this representation we see that: between any two knots, the spline is a cubic function; the spline and its first two derivatives are continuous; and the spline is linear for $t < t_1$ and $t > t_n$. For our purposes, the most important property of the spline is the following. Given data $(t_1, y_1), ..., (t_n, y_n)$, and $s(t)$ a differentiable function with a continuous first derivative, the minimum value of

$$\sum [y_i - s(t_i)]^2 + \int \{s''(t)\}^2 dt$$

is attained when *s(t)* is a NCS. The second term is a roughness penalty and the solution to this minimization problem is termed a smoothing spline.

Mixed model formulation

For a general spline, the values at the knots are

$$s = T\Phi + Dc$$

where $\{t_i^{j-1}\}$, $D = \frac{1}{6}\{(t_i - t_j)_+^3\}$, $\Phi^T = (\Phi_0, \Phi_1)$, and $c^T = (c_1, \ldots, c_n)$. Define Q $(n \times n - 2)$ to be the divided difference matrix with entries

$$q_{j,j} = h_j^{-1},\ q_{j+1,j} = -h_j^{-1} - h_{j+1}^{-1},\ q_{j+2,j} = h_{j+1}^{-1}$$

where $h_j = t_{j+1} - t_j$ and $q_{ij} = 0$ for $i > j + 2$. Since $T' Q = 0$, one way to allow for the constraints $Tc = 0$ is to write $c = Q\gamma$ for an unconstrained vector γ of dimension $n - 2$. Then

$$s = T\Phi + L\gamma$$

where $L = DQ$. Solving $y = T\Phi + L\gamma$ (n equations for n unknowns) produces the equation of the interpolating spline which passes exactly through the data points. With this formulation, $\gamma_1 \ldots \gamma_{n-2}$ are the values of the second derivative of *s(t)* at the internal knots $t_2 \ldots t_{n-1}$, and the roughness penalty is $\gamma^T R\gamma$, where R is a symmetric matrix of dimension $n - 2$ with elements

$$r_{i,i} = \tfrac{1}{3}(h_i + h_{i+1}),\ r_{i,i+1} = r_{i+1,i} = \tfrac{1}{6}h_{i+1}$$

and $r_{i,j} = 0$ when $|i - j| > 1$. The smoothing spline minimizes

$$(y - T\Phi - L\gamma)^T(y - T\Phi - L\gamma) + \alpha\gamma^T R\gamma$$

as does the BLUP solution to the mixed model with fixed effects $T\Phi$, random effects $L\gamma$, and cov(γ) % R^{-1} (Robinson, 1991). Thus splines can be fitted by BLUP (or REML) calculations.

The mixed model equations are

$$(T'T)\Phi + (T'L)\gamma = T'y$$
$$(L'T)\Phi + (L'L + \alpha R)\gamma = L'y$$

Solving in the usual way produces

$$\hat{\Phi} = (T'T)^{-1}T'(y - L\hat{\gamma})$$

and

$$\{\alpha R + L'[I - T(T'T)^{-1}T']L\}\hat{\gamma} = L'[I - T(T'T)^{-1}T']y$$

Using (a) $I - T(T' T)^{-1}T' = Q(Q' Q)^{-1}Q'$ and (b) $L' Q = Q' L = R$ allows the second equation to be simplified to

$$(R + \alpha Q'Q)\hat{\gamma} = Q'y$$

the well-known Reinsch equations for the smoothing spline (Reinsch, 1967). The fitted values can be expressed as

$$s = T\hat{\Phi} + L\hat{\gamma} = T(T'T)^{-1}T'y - T(T'T)^{-1}T'L\hat{\gamma} + L\hat{\gamma}$$

The first term represents the linear regression of y on t. Again using the results (a) and (b), the next two terms can be reduced to

$$[I - T(T'T)^{-1}T']L\hat{\gamma} = Q(Q'Q)^{-1}Q'L\hat{\gamma} = Q(Q'Q)^{-1}R\hat{\gamma}$$

The spline can thus be represented as deviations from the least-squares regression line by setting the random component of the model to Zu, with $Z = Q(Q' Q)^{-1}R$ and $u = \gamma$. A further simplification is to make the covariance matrix of u proportional to the identity matrix, e.g., by choosing $u = P^T\gamma$ and $Z = Q(Q' Q)^{-1}P$, where P is the Cholesky root of R.

Example

ASREML (Gilmour *et al.,* 1997) was used to fit splines to first lactation test day records for the progeny of 535 sires, a total of 2351 cows in 140

herds. The hierarchical model had linear regression and spline terms for the general mean, sires, and cows within sires. The regression coefficients (intercept and slope) were treated as random and possibly correlated. In order to keep the number of spline knots low, the covariate 'days in milk' was grouped by test day and coded 1 ... 10. The analysis adjusted for the additive effects of herd and test-month.

Results and discussion

Table 1 shows the estimated components of variance and covariance. The spline terms, which were highly significant, play a role similar to the exponential term of Wilmink's curve (Wilmink, 1987), or the higher order terms in a random regression model, but in a more flexible manner. It should be relatively straightforward to combine the estimated variance components with the various matrices *(Q,R,* etc.) presented above to produce estimates of genetic and environmental covariance matrices, and hence heritabilities. This work is in progress. Breeding values for the sires are available from the fitted values (regression + spline) at the sire level, and the sum of these over all ten test days, suitably scaled, provides an estimate of 305-day yield. These were found to be in reasonably good agreement with national evaluations.

In general the spline factor can be combined with other terms in the model to produce interactions with fixed effects or nesting within other random effects such as sire or animal. In a sire analysis, animal#spline models the environmental variation (rather like an animal#trait term in a multivariate analysis) and sire#spline (equivalent to sire#trait) models the genetic effects. The model can easily accommodate any additional fixed effects expected to affect the lactation curve differently at different stages. For example, age at calving might be included as an age#spline term. When the tests are at approximate 30-day intervals, but the cows enter the first test with different numbers of days in milk, M says, this can be allowed for by including a M#Spline interaction term in the model. This spline should look like the first derivative of a lactation curve.

Table 1. Variance components for regression and spline terms.

	Regression			
	Intercept	Slope	Covariance	Spline
Mean				1.62549
Sire	1.52195	0.02479	-0.13465	0.20757
Cow (sire)	11.42590	0.15255	-0.92907	0.35290
Residual	3.73973			

References

Gilmour, A.R., Thompson, R., Cullis, B.R. and Welham, S.J. 1997. *The AS-REML Manual.*

Robinson, G.K. 1991. That BLUP is a good thing: the estimation of random effects. *Statistical Science 6,* 15-51.

Reinsch, C. 1967. Smoothing by spline functions. *Numer. Math. 10,* 177-183.

Verbyla, A.P., Cullis, B.R., Kenward, M.G. and Welham, S.J. 1997. Smoothing splines in the analysis of designed experiments and longitudinal data. (To appear.)

Wilmink, J.B.M. 1987. Adjustment of test-day milk, fat and protein yields for age, season and stage of lactation. *Livest. Prod. Sci. 16,* 335-348.