STATISTICAL MODELS FOR THE GENETIC ANALYSIS OF LONGITUDINAL DATA

F. Jaffrézic[1,2], I.M.S. White[2], R. Thompson[3,4], P.M. Visscher[2] and W.G. Hill[2]

[1] INRA, SGQA, 78352 Jouy-en-Josas, France
[2] Institute of Cell, Animal and Population Biology, University of Edinburgh, UK
[3] Rothamsted Experimental Station, IACR, Harpenden, UK
[4] Roslin Institute (Edinburgh), Roslin, Midlothian, UK

INTRODUCTION

A simple and efficient procedure for the genetic analysis of characters that change as a function of age (or some other independent and continuous variable) is desirable. Three methodologies have been put forward in the literature: random regression, character processes and structured antedependence models. The aim of this paper is to compare the different approaches for the genetic analysis of longitudinal data. We focus on examining the underlying assumptions of the three models, on describing the types of covariance structures (genetic, environmental) accommodated by each method and on evaluating their ability to adequately fit empirical data.

MODELS

Genetic Analysis. As with traditional quantitative genetics, it is assumed that the observed phenotypic trajectory can be decomposed as: $Y(t) = \mu(t) + g(t) + e(t) + \varepsilon$, where $\mu(t)$ is a nonrandom function, the genotypic mean function of *Y(t),* ε is the residual variation which is assumed normally distributed with unknown variance, *g(t)* and *e(t)* are Gaussian random functions and represent the age-dependent genetic and environmental deviations, respectively. They are independent of one another, of mean zero, with covariance functions *G(s,t)* and *E(s,t).* Three different methodologies are presented to model these covariance functions.

Random regression models. Random regression (RR) models are already well known in the context of longitudinal data analysis (Diggle *et al*., 1994), and we will not give a detailed description of these models here. The primary objective is to choose the most appropriate parametric functions for individual deviations in the genetic and permanent environmental parts. Genetic and environmental covariances as a function of age are then determined by the variances and covariances among the regression coefficients. It has been shown (Meyer and Hill, 1997) that orthogonal polynomial (OP) models, initially proposed by Kirkpatrick and Heckman (1989) as a non-parametric way of smoothing previously estimated covariance matrices, can be considered as a special case of random regression models.

Character process models. In contrast to random regression, the character process model does not attempt to model the forms of the *g(t)* or *e(t)* functions. Instead, parametric models for the covariance functions themselves (i.e., *G(s,t)* and *E(s,t)*) are the target of analysis (Pletcher and Geyer, 1999). Taking the genetic covariance function as an example, the covariance function can be decomposed into $G(s,t) = v_G(s)v_G(t)\rho_G(|s-t|)$ where $v_G(t)^2$ describes how the

genetic variance changes with age and $\rho_G(|s-t|)$ describes the genetic correlation between two ages. There are no restrictions on the form of $v_G(.)$, and it is often modeled using simple polynomials (linear, quadratic, etc.). As presented by Pletcher and Geyer (1999), correlation-stationarity was assumed, i.e. the correlation between two ages is assumed to be a function only of the time distance (|*s*-*t*|) between them. We proposed (Jaffrézic and Pletcher, 2000) an extension of the character process model for non-stationary correlations using a methodology suggested by Nunez-Anton and Zimmerman (2000). The basic idea is to implement a non-linear transformation upon the time axis, *f(t),* such that correlation stationarity holds on the transformed scale; on the original scale the correlation is non-stationary. The correlation function is then defined as $\rho(s,t) = \rho(|f(s)-f(t)|)$, and the functions suggested by Pletcher and Geyer (1999) remain valid. Ideally the transformation function should contain a small number of parameters with interpretable effects.

Structured antedependence models. The basic idea of antedependence models is that an observation at time *t* can be explained by the previous ones. An antedependence structure of order *r* is defined by the fact that the *i*th observation (*i* > *r*) given the *r* preceding ones is independent of all previous observations (Gabriel, 1962). Generalizing this concept to genetic analysis, a second order structured antedependence model for the genetic part *g(t)* can be written as:

$$g(t_0) = \varepsilon_g(t_0)$$

$$g(t_1) = \phi_1 g(t_0) + \varepsilon_g(t_1)$$

$$g(t_j) = \phi_1 g(t_{j-1}) + \phi_2 g(t_{j-2}) + \varepsilon_g(t_j)$$

for j ≥ 2. Here, ϕ_1 and ϕ_2 are regression parameters, and $\varepsilon_g(t)$ is assumed to be normally distributed, with mean zero and variance $\sigma_g^2(t)$ that can change with time. This corresponds to a generalization of simple autoregressive models that assume constant variances. In structured antedependence (SAD) models, Nunez-Anton and Zimmerman (2000) propose a parametric function for innovation variances $\sigma_g^2(t)$ using for example a polynomial of time. SAD models require very few parameters for the covariance structure, and increasing the order of antedependence only involves one extra parameter at each step. The same model can be written for environmental effects *e(t)*.

COMPARISONS FOR UNIVARIATE ANALYSIS

All the analyses were performed using the software ASREML (Gilmour *et al*., 2000). Through extensive investigation of a variety of simulated covariance structures and empirical data, it was found that under most conditions character processes and structured antedependence models provide the best description of the underlying covariance structure. CP models could in particular deal very well with a correlation function decreasing asymptotically to zero which was not well fitted by random regression models. In fact, as shown on Figure 1, as polynomials do not have asymptotes, the estimated correlation goes negative. This is an important drawback in practice as this kind of asymptotic pattern can often be expected. A further advantage of the

CP models appears to be the ability to model the variance and correlation separately. As mentioned previously, for random regression models the entire covariance structure is implicity determined by the shapes of the regression polynomials, and covariance surfaces described by orthogonal polynomials have a fixed relationship between variance and correlation. It is also likely that the separation of variance and correlation was a major factor contributing to the ability of the CP model to estimate the genetic variation with a much smaller number of parameters than random regression. Comparisons performed by Meuwissen and Pool (2001) also showed that CP models have good prediction properties.

SAD models appeared to have similar advantages to CP models in fitting the covariance structure with few parameters and were also able to deal with the highly non-stationary environmental correlation pattern found in a milk production analysis for dairy cattle based on test day records. The analyses were performed on a data set that comprised 9277 cows from 464 bulls with 10 records per animal (F. Jaffrézic, PhD thesis). A sire model was considered here - the environmental covariance structure in an animal model would be simpler and easier to fit. As shown in Table 1, it appeared that a structured antedependence model of first order for the genetic part and third order for the environmental part (Model 3) had a higher likelihood than a quartic random regression model (Model 8) with far fewer parameters (11 instead of 31).

DISCUSSION

These analyses showed that SAD models offer similar advantages to character processes over random regression to fit the covariance structure with few parameters. They also proved to be able to deal with quite complex non-stationary correlation patterns. Another advantage of these models is that they offer an extension to multi-trait analysis, which is not yet possible for CP models. Extension of RR models to the multivariate case is straightforward. It does, however, require a large number of parameters. For example, a bivariate genetic analysis considering only a quadratic RR for both genetic and environmental parts requires 45 parameters. When increasing to the cubic order for both parts, the number of parameters jumps to 75. In contrast, increasing the order of structured antedependence model adds only 2 parameters at each step. Analyses on a variety of simulated and real data sets showed that despite their much smaller number of parameters, bivariate SAD models offer a high degree of flexibility to fit the cross-covariance structure and, in general, perform better than random regression models.

Some questions are still open for the design and utility of SAD models. In particular, parametric forms for variance and correlation functions are, in general, difficult to obtain and their relationship with regression parameters ϕ and innovation variances should be further studied. Further flexibility could also be achieved in the SAD models by considering heterogeneous variances as suggested by Foulley and Quaas(1995) or by allowing the regression coefficients ϕ to change with time as suggested by Nunez-Anton and Zimmerman (2000). Structured antedependence models therefore offer a very promising direction for the genetic analysis of longitudinal data and are likely to attract increasing attention in the near future.

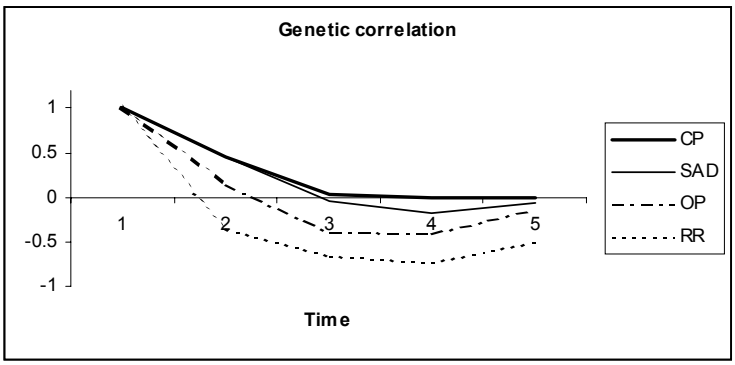

Figure 1 : Genetic correlation simulated according to a stationary exponential CP model and estimated with a linear random regression (RR), cubic orthogonal polynomial (OP) and second order antedependence model with linear innovation variance (SAD)

Table 1 : Model comparisons for the genetic analysis of lactation curve for dairy cattle (NPCov : number of parameters in the covariance structure, LogL : Log-likelihood, US : unstructured covariance matrix, SAD(*i*) : structured antedependence model of order *i*)

Model	Genetic	Environmental	NPCov	LogL
Unstructured				
1	US	US	110	4126
2	SAD(1)	US	59	4109
Structured Antedependent				
3	SAD(1)	SAD(3)	11	3845
4	SAD(2)	SAD(3)	12	3852
5	SAD(3)	SAD(3)	13	3854
6	SAD(2)	SAD(2)	11	3796
7	SAD(1)	SAD(1)	9	3580
Random Regression				
8	Quartic	Quartic	31	3623
9	Quadratic	Quartic	22	3607
10	Cubic	Cubic	21	3336
11	Quadratic	Quadratic	13	2767

REFERENCES

Diggle, P.J. *et al*. (1994) "Analysis of Longitudinal Data". Oxford, Clarendon Press.
Foulley, J.L. and Quaas, R.L. (1995) *Genet. Sel. Evol.* **27** : 211-228.
Gabriel, K.R. (1962) *Ann. Math. Stat*. **33** : 201-212.
Gilmour, A.R., Thompson, R., Cullis, B.R. and Welham, S.J. (2000) "ASREML Manual". New South Wales Department of Agriculture, Orange, Australia.
Jaffrézic, F. and Pletcher, S.D. (2000) *Genetics* **156** : 913-922.
Kirkpatrick, M. and Heckman, N. (1989) *J. Math. Biol.* **27** : 429-450.
Meuwissen, T.H.E. and Pool, M.H. (2001) *Interbull Bulletin.* **27** : 172-178.
Meyer, K. and Hill, W.G. (1997) *Livest. Prod. Sci.* **47** : 185-200.
Nunez-Anton, V. and Zimmerman, D.L. (2000) *Biometrics* **56** : 699-705.
Pletcher, S.D. and Geyer, C.J. (1999) *Genetics* **153** : 825-833.