



Variance component estimation with longitudinal data: a simulation study with alternative methods

Simone Inoe Araujo^{1*}, Adair José Regazzi², Claudio Vieira de Araujo¹, Cosme Damião Cruz², Carlos Henrique Osório Silva², and José Marcelo Soriano Viana²

Received 10 March 2008

Accepted 24 March 2009

ABSTRACT - A pedigree structure distributed in three different places was generated. For each offspring, phenotypic information was generated for five different ages (12, 30, 48, 66 and 84 months). The data file was simulated allowing some information to be lost (10, 20, 30 and 40%) by a random process and by selecting the ones with lower phenotypic values, representing the selection effect. Three alternative analysis were used, the repeatability model, random regression model and multiple-trait model. Random regression showed to be more adequate to continually describe the covariance structure of growth over time than single-trait and repeatability models, when the assumption of a correlation between successive measurements in the same individual was different from one another. Without selection, random regression and multiple-trait models were very similar.

Key words: Random regression, multiple-trait, repeatability, selection.

INTRODUCTION

Data obtained by successive measurements in one experimental unit or individual, the so called longitudinal data, can be analyzed using different strategies. In genetic breeding an interesting approach is to use random regression. Henderson Junior (1982) proposed the theory regarding the random regression coefficients, based on the principle that if a regression coefficient pertaining to each individual in an experiment is defined, and if the individuals are a random sample of the population, then the regression coefficients must be considered random. This methodology has been used to model traits that are measured over time, such as growth traits. In order to adopt this model, the measurements over time are considered to be successive points on a continuous trajectory, and hence a prediction

of parameters is permitted, also for points (ages) where measurements have not been done. To describe the fixed curve for all the individuals, as well as the individual ones, covariance functions (Kirkpatrick et al. 1990) that describe the covariance structure between ages can be used. In this context, covariance functions using Legendre's polynomials have been used because they make calculations and interpretations easier.

Another alternative of analysis for longitudinal data is the repeatability model, but it is necessary to assume that successive measurements in the same individual with the present correlation equal to the unit. However, such assumptions are not always valid ones, because for growth traits, for example, successive measurements are always more strongly correlated than those more distant in time.

¹ Universidade Federal de Mato Grosso, Avenida Alexandre Ferronato, 78550-000, Sinop, MT, Brazil. *E-mail: araujoine@bol.com.br

² Universidade Federal de Viçosa, 36571-000, Viçosa, MG, Brazil

In a multi-trait approach we do not need to make assumptions regarding the covariance structure, which means that a non structured covariance matrix is used. Consequently, this alternative requires the estimation of a large number of parameters, and may cause many computational difficulties.

The objective of this paper was to study variance components and genetic parameter estimations using random regression, repeatability and the multi-trait models with longitudinal data with different levels of lost information present in the data.

MATERIAL AND METHODS

The data was simulated such that it represented crossing structures distributed in three different places. A progeny test was simulated, by crossing 30 males each with three different females, with each crossing generating ten different offspring. This crossing structure was simulated in three different places. The fixed effect of place was created to present non significant statistical differences (similar means and variances). For each offspring there were phenotypic data at five different ages, 12, 30, 48, 66 and 84 months, resulting in 120 relatives (30 males and 90 females), giving a total of 1,020 individuals, with 900 of them presenting information at five different ages, which resulted in 4,500 observations of production.

The simulated longitudinal data can be described by the mixed linear model defined as: $y = X\beta + Za + Wp + \epsilon$, where y is the vector of observations from each individual; β represents the vector of fixed effect and of the general curve parameters for all the individuals; X is the incidence matrix of fixed-effects levels and the regression variables, corresponding to the standardized ages associated to the Legendre's polynomials; a and p are the random vectors of random regression solutions of the additive genetic effect and of random regression of the permanent environmental effects, respectively; Z and W are the matrices that associate the standardized ages by the Legendre's Polynomials to the a and p vectors, respectively. ϵ is the random temporary environmental effect vector.

Assuming that the vectors y , a , p e ϵ have a normal distribution then $E(a) = 0$, and it's variance is $V(a) = A \otimes Ka = G$, where Ka is a covariance matrix between the random regression coefficients of the additive genetic effect and A is a matrix that indicates

the degree of the individuals relationship, of equal dimension to the total number of individuals (N). The vector p has $E(p) = 0$ and a variance $V(p) = I \otimes Kp = P$, where Kp is a covariance matrix between random regression coefficients of the permanent environmental effect and I is an identity matrix of equal dimensions to the number of individuals with information (n). Finally, ϵ has the expected value $E(\epsilon) = 0$, and a variance of $V(\epsilon) = I\sigma_e^2 = R$, where σ_e^2 is the temporary environmental effect variance. Consequently, y has a mean and variance equal to $E(Y) = X\beta$ and $V(Y) = ZGZ' + WPW' + R$, respectively.

The fixed effects and the general regression curve for all the individuals are associated to the b vector. The data simulation was carried out by using a second degree polynomial model, with Legendre's orthogonal polynomials to describe both the fixed trajectory and the random effects of the model. The temporary residual effect was assumed to have a normal distribution with an average equal to zero and a variance of $\sigma_e^2 = 2.2$ unity² and the Ka and Kp covariance matrices of random regression coefficients of the additive genetic and permanent environmental effect, were respectively defined as follows,

$$Ka = \begin{bmatrix} 2.2367 & 0.5571 & -0.1227 \\ 0.5571 & 0.2939 & 0.0027 \\ -0.1227 & 0.0027 & 0.0328 \end{bmatrix} \text{ and}$$

$$Kp = \begin{bmatrix} 1.4140 & 0.1405 & -0.3454 \\ 0.1405 & 0.5960 & 0.1328 \\ -0.3454 & 0.1328 & 0.4869 \end{bmatrix}$$

The β vector containing the fixed curve was obtained by $\beta = (\phi' \phi)^{-1} (\phi' \gamma)$, where $\gamma' = [15.14 \ 26.82 \ 35.72 \ 41.84 \ 45.1]$ represents the average production at months 12, 30, 48, 66 and 84, respectively. The ϕ matrix represents the multiplication between the standardized ages matrix (M) with the one that describes the three first Legendre's polynomials (Λ), $\phi = M\Lambda$, resulting in $\beta' = (\phi' \phi)^{-1} (\phi' \gamma)' = [47.8832 \ 12.2381 \ -2.3636]$, which is the vector of solutions for the fixed curve for all the individuals in the study representing the intercept and the linear and quadratic coefficients of the equation, respectively. If $UT_a^{1/2}$ and $UT_p^{1/2}$ are the Cholesky decomposition of the covariance matrices of the random regression coefficients for additive genetic and permanent environmental effects, respectively, and

A is a numerator relationship matrix between individuals, where $AT^{1/2}_a$ is the Cholesky decomposition of this relationship matrix, then the vector y (phenotypes) containing the "i"th traits (age) is defined as $y = \beta + AT^{1/2}_a Z a UT^{1/2}_a + Z p UT^{1/2}_p + e$.

After the complete data file simulation 450 individuals with information were randomly chosen, which had the production information at 84 months deleted giving a second file with a 10% loss of information. Subsequently, the same individuals had deleted the production information at months 66, 48 and 30, giving new data files with 20%, 30% and 40% losses of information, respectively. The elimination of information from individuals using a random process aimed at studying the efficiency of the common methods used in longitudinal data analysis, when applied to incomplete data. Our goal was to study the lost observations effects.

Again, with the complete data file, the individual elimination process was made. However in this case, the individuals whose information was deleted were the ones who had the least phenotypic value and represented samples with the selection effect, with the objective to study the efficiency of the common strategies analysis used in longitudinal data analysis, regarding the effect of lost observations by the selection effect.

Production data in each situation were analyzed to estimate the (co)variance components and the genetic parameters by different methods. Both the simulation and manipulation of the data files were realized with the Statistical Analysis System version 8.1 (SAS 1990, SAS Institute Inc.).

Considering an analysis where each age is like a repeated measurement in the same individual, the repeatability model is described as $y = Xb + Za + Wp + e$, where y is a $n \times 1$ vector of n observations (production), X is an incidence matrix of local fixed effects and of the age (co) variable for each individual's production, Z and W are the incidence matrices of the individual's random additive genetic and permanent environmental effects associated to vectors a and p of the additive genetic and permanent environmental values, and e is the residual vector with the same dimension of y .

The random regression model used considered each age as a point in a continuous trajectory, fitting covariance functions both for the additive genetic effect and for the permanent environmental one, where both

functions used the first three Legendre's polynomials, characterizing a second degree polynomial function. This model can be described as $y = X\beta + Za + Wp + e$; where y is the vector of n observations of production at each age; X is the incidence matrix of local fixed effects and the standardized ages between -1 to +1 that describes the averaged trajectory of all individuals by the Legendre's polynomials; β is the solutions vector of local fixed effects and of the fixed regression of all the individuals; Z and W are diagonal block matrices with the standardized ages associated to the random regression coefficients of the additive genetic and permanent environmental random effects for each individual, respectively; a and p are random regression model vectors of additive genetic and permanent environmental effects, respectively, for each individual. The vector e represents the random temporary environmental effects.

In the multiple trait analysis each age was considered a distinct characteristic. The model used is described as $y = X\beta + Za + e$, where :

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}; X = \begin{bmatrix} X_1 & \phi \\ \phi & X_2 \end{bmatrix}; \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}; Z = \begin{bmatrix} Z_1 & \phi \\ \phi & Z_2 \end{bmatrix};$$

$$a = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}; e = \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}.$$

y_i is the response variable vector; X_i is the incidence matrix of the local fixed effects; β is the local fixed effects levels solutions; Z is an incidence matrix of random effects; a and e are the additive genetic and residual random effect vectors, respectively, at the i 'th age.

$E[y] = X\beta$ and $V[y] = ZGZ' + R$ where $G = A \otimes G_0$ is the genetic additive variance and covariance matrix and $R = I \otimes R_0$ where R_0 is the residual variance and covariance matrix.

To compare the results from different data files with different strategies of analysis, genetic parameters and variances estimates were evaluated. In some cases the likelihood ratio test was used (Rao 1973).

The analyses were all processed with the software DFREML Version 3.0 (Meyer 1998).

RESULTS AND DISCUSSION

With the repeatability model analysis and random loss of information, the heritability estimates were similar

between the different ages (Table 1). However, when the loss of information occurred by selection, there was a reduction in the additive genetic variance, according to the loss of information level.

The additive genetic variance estimates at each age, for the complete data file with loss of information, with and without selection, by using random regression models are presented in Table 2.

The loss of information obtained with the random process did not modify the heritability estimates at any loss of information level (Table 3). According to Dal Zotto (2000), in random regression models the minimum number of observations to be considered for each permanent environmental effect level must be equal to the number of parameters used to describe the data trajectory plus one. According to Schaeffer and Dekkers (1994), random regression models can also be used when individuals have only one piece of information. Results obtained in this study are in agreement with these authors.

When the loss of information was made by eliminating the individuals with the least phenotypic values, the genetic and environmental variance estimates reduced, causing smaller values for the heritability estimates. With selection, only the random regression heritability estimates with a 10% loss of

information were similar to the complete data file analysis estimates. Resende et al. (2001), using random regression models to describe the diameter at breast height within one to seven year old *Eucalyptus urophylla* trees, found that the heritability estimates were very similar to that obtained by single-trait models before three years old. For advanced ages the estimates were smaller. The authors discussed that single-trait analysis overestimated the parameter estimates because of the phenotypic variance reduction, due to the least vigorous individual death, became naturally selected for population adaptation. Matheson and Raymond (1984), cited by Resende et al. (2001), found heritability estimates in two *Pinus radiata* populations equal to 0.12 and 0.24. The estimates assumed values of 0.21 and 0.33 after the worst plants were eliminated, respectively. These results were opposite to those obtained in this study, probably due to the sample size, since our data was simulated considering a progeny test with only two generations; the ancestry and the offspring. Therefore, the elimination of information from smaller phenotypic values and consequently the elimination of the smaller genotypic values changed the data structure.

Table 1. Additive genetic variance, permanent environmental variance, temporary environmental variance, environmental variance, and heritability estimates, by using the repeatability model, with and without selection

Estimates	Loss of information without selection (%)					Loss of information with selection (%)			
	0	10	20	30	40	10	20	30	40
$\hat{\sigma}_a^2$	1.225	1.101	0.994	0.969	0.979	0.970	0.694	0.464	0.299
$\hat{\sigma}_p^2$	0.217	0.271	0.285	0.280	0.311	0.215	0.121	0.058	0.012
$\hat{\sigma}_e^2$	3.666	3.358	3.401	3.398	3.481	3.368	3.357	3.325	3.378
$\hat{\sigma}_p^2 + \hat{\sigma}_e^2$	3.883	3.629	3.686	3.678	3.792	3.583	3.478	3.383	3.391
\hat{h}^2	0.24	0.23	0.21	0.21	0.21	0.21	0.17	0.12	0.08

Table 2. Estimates of additive genetic variance in each age for the complete and incomplete data file, with and without selection, by using random regression models

Age (Months)	Data Complete	Loss of information without selection (%)				Loss of information with selection (%)			
		10	20	30	40	10	20	30	40
12	0.613	0.606	0.613	0.569	0.530	0.596	0.599	0.667	0.580
30	0.683	0.719	0.719	0.827	0.901	0.734	0.789	0.695	0.313
48	1.010	1.080	1.044	1.196	1.320	1.028	0.785	0.480	0.149
66	1.793	1.920	1.818	1.845	1.933	1.683	0.693	0.309	0.176
84	3.304	3.549	3.359	3.031	3.030	2.987	0.784	1.015	0.932

Table 3. Estimates of heritability in each age for the complete and incomplete data file, with and without selection, by using random regression model

Age (Months)	Complete Data	Loss of information without selection (%)				Loss of information with selection (%)			
		10	20	30	40	10	20	30	40
12	0.16	0.15	0.15	0.15	0.14	0.15	0.15	0.17	0.15
30	0.18	0.18	0.18	0.21	0.23	0.19	0.19	0.19	0.09
48	0.22	0.23	0.22	0.25	0.28	0.22	0.18	0.13	0.04
66	0.35	0.38	0.36	0.37	0.39	0.35	0.21	0.11	0.06
84	0.41	0.46	0.44	0.40	0.41	0.43	0.16	0.20	0.19

The complete data files were analyzed by two random regression models, where the difference between them was in the temporary environmental effect variance. In the first model the variances were constant and in the second model each age had a different variance. Comparing the estimates of variance components and heritability, and also comparing the models likely function values (Table 4), its acceptance at a 5% significance level indicates that the temporary environmental variances being homogeneous or heterogeneous did not influence the model fit.

In the data analysis considering each age as a trait by multi-trait models analyzing traits side by side (bi-trait model) the same results were observed. The heritability estimates for the complete and incomplete data files with and without selection are presented in Table 5. These values represent the minimum and maximum estimates obtained by the analysis of traits side by side. The selection caused a variability reduction since the loss of information increased, in other words this model was very sensitive to the selection effect. However, all the heritability estimates were similar when the loss of information was not made by a selective process.

Generally, independent from the analysis strategy used (single-trait, multi-trait or random regression models) the selection effect caused the least

discrimination among individuals and reduced the variability between them. The genetic variance was reduced as a result of less variability between individuals and consequently smaller heritability estimates. With a 10% loss of information the heritability values did not change, with a 20% loss the heritability values reduced at the last two ages, with a 30% loss the reduction in heritability values occurred at the three last ages and with a 40% loss the reduction occurred at the four last ages. However with a 10% loss of information with selection the random regression model analysis was the best alternative.

Knowing that the loss of information could change the temporary environmental effects, the data was also analyzed with a 40% loss of information with selection, but assuming heterogeneous variance for the temporary environmental effect at each age. Moreover the complete data file was also analyzed to check the permanent environmental effect of variance homogeneity effects. Comparing the variance components and heritability estimates and comparing the models using the likelihood function values (Table 6) it was observed that the variance heterogeneity for both the complete and the select data file did not change the data variation description. Therefore the loss of information did not change the temporary environmental effect.

Table 4. Additive genetic variance (σ_a^2), permanent environment (σ_p^2), and temporary environment (σ_e^2), and heritability (h^2) estimates, in each age by random regression models with different assumptions, likelihood functions and likelihood ratio tests (χ^2) values

Ages	Homogeneous residual				Heterogeneous residual			
	σ_a^2	σ_p^2	σ_e^2	h^2	σ_a^2	σ_p^2	σ_e^2	h^2
12	0.623	1.236	2.067	0.16	0.586	1.263	2.043	0.15
30	0.705	1.048	2.067	0.18	0.707	1.031	2.084	0.18
48	1.089	1.537	2.067	0.23	1.116	1.532	2.113	0.23
60	1.811	1.223	2.067	0.35	1.876	1.198	1.918	0.38
84	2.916	2.864	2.067	0.37	3.083	2.523	2.389	0.39
-2Log(L)	10624.23				10621.64			
$\chi^2 2.59$ ($P > 0.05$)								

Table 5. Heritability estimates using the complete and incomplete data file considering or not the selection by the multi-trait model in the uni-trait analysis

Ages	Complete	Loss of information without selection (%)			
		10 %	20 %	30 %	40 %
12	0.12–0.14	0.13–0.14	0.13–0.14	0.11–0.13	0.12–0.13
30	0.20–0.21	0.20–0.25	0.20–0.25	0.20–0.25	0.27–0.29
48	0.19–0.23	0.19–0.23	0.19–0.23	0.20–0.26	0.23–0.26
66	0.39–0.42	0.33–0.40	0.38–0.42	0.38–0.42	0.38–0.40
84	0.37–0.40	0.32–0.43	0.38–0.41	0.38–0.41	0.38–0.40

Ages	Complete	Loss of information with selection (%)			
		10 %	20 %	30 %	40 %
12	0.12–0.14	0.13–0.14	0.13–0.15	0.13	0.13
30	0.20–0.21	0.20–0.25	0.19–0.20	0.20–0.25	0.12–0.14
48	0.19–0.23	0.19–0.23	0.19–0.22	0.05–0.07	0.05–0.06
66	0.39–0.42	0.40–0.42	0.03–0.11	0.03–0.07	0.03–0.06
84	0.37–0.40	0.11–0.23	0.11–0.18	0.15–0.19	0.16–0.18

Table 6. Additive genetic ($\hat{\sigma}_a^2$), permanent environment ($\hat{\sigma}_p^2$), temporary environment ($\hat{\sigma}_e^2$) variances, heritability (\hat{h}^2) estimates, likelihood function logarithm and likelihood ratio test (χ^2) values for the complete and selected data with a 40% loss of information in random regression models

Complete data		Homogeneous residual				Heterogeneous residual			
Ages	$\hat{\sigma}_a^2$	$\hat{\sigma}_p^2$	$\hat{\sigma}_e^2$	\hat{h}^2	$\hat{\sigma}_a^2$	$\hat{\sigma}_p^2$	$\hat{\sigma}_e^2$	\hat{h}^2	
12	0.613	1.252	2.062	0.16	0.618	1.261	2.044	0.16	
30	0.683	1.079	2.062	0.18	0.702	1.049	2.090	0.18	
48	1.010	1.598	2.062	0.22	1.044	1.561	2.154	0.22	
60	1.793	1.228	2.062	0.35	1.825	1.187	1.890	0.37	
84	3.304	2.630	2.062	0.41	3.277	2.337	2.480	0.40	
-2Log(L)		10624.22				10621.64			
χ^2		2.58 (P > 0.05)							

Selection data		Homogeneous residual				Heterogeneous residual			
Ages	$\hat{\sigma}_a^2$	$\hat{\sigma}_p^2$	$\hat{\sigma}_e^2$	\hat{h}^2	$\hat{\sigma}_a^2$	$\hat{\sigma}_p^2$	$\hat{\sigma}_e^2$	\hat{h}^2	
12	0.580	1.383	1.984	0.15	0.565	1.509	1.838	0.14	
30	0.313	1.009	1.984	0.09	0.298	0.993	2.230	0.08	
48	0.149	1.214	1.984	0.04	0.141	1.215	1.963	0.04	
60	0.176	0.545	1.984	0.06	0.184	0.517	1.693	0.07	
84	0.932	1.906	1.984	0.19	0.920	1.495	2.588	0.18	
-2Log(L)		5979.28				5970.94			
χ^2		8.34 (P > 0.05)							

The heritability estimates arising from random regression models when using complete or incomplete data files with and without selection were similar to the multi-trait models analysis estimates. However the heritability estimates obtained with the multi-trait model

were more affected by the loss of information data than those obtained by random regression models. With a 10% loss of information the heritability estimates obtained by the random regression model were similar to those obtained with the complete data file. However

the multi-trait model at the same loss of information level presented underestimated values at the last age studied. Therefore as the random regression models had used covariance functions that give a continuous covariance structure of random effects associated to the analyzed character description, it can be affirmed that they can better express the mixed linear model's random effects variance that describe longitudinal data. By considering that the trait in study can change in time, this strategy is more realistic than the repeatability model, and because it uses fewer parameters it becomes more attractive than the multi-trait models that can be prohibited in practical applications with a large number of parameters.

Schaeffer and Wilton (1998) discussed that in some cases when all the traits are observed in each individual and the traits heritability are similar and all of them are positively correlated, the model analysis that considers multiple traits would not offer a significant increase in the genetic evaluation accuracy.

CONCLUSIONS

When the assumption is that the correlation between successive measurements in a single individual is equal to one which is invalid, then random regression models in the genetic analysis can provide better results than repeatability models.

Random regression models yield covariance and genetic parameter estimates similar to those obtained by multiple-trait models, however, with the use of fewer parameters in the model, which in practice is an advantage.

In small population samples submitted to the selection effect, multi-trait models were more susceptible to the selection bias than the random regression models analysis, but additional studies are necessary for more conclusive findings.

Considering the continuous nature of the dependent variable, random regression models must be preferred to single trait models in high loss of information level data files.

Estimação de componentes de variância com dados longitudinais: um estudo de simulação com métodos alternativos

RESUMO - Foi gerada uma estrutura de pedigree, considerando-se três locais diferentes. Para cada prole foram geradas informações fenotípicas em cinco idades diferentes (aos 12, 30, 48, 66 e 84 meses). O arquivo de dados foi gerado permitindo que algumas informações fossem perdidas (10, 20, 30 e 40%) por um processo aleatório e por aqueles indivíduos com menor valor fenotípico, representando o efeito da seleção. Foram utilizados Modelos de repetibilidade, de regressão aleatória e multi-característica. Modelos de regressão aleatória foram mais adequados para descrever continuamente as estruturas de covariâncias de crescimento ao longo do tempo, do que modelos de repetibilidade, quando a pressuposição de que a correlação entre mensurações sucessivas no mesmo indivíduo é diferente da unidade. Sob ausência de seleção, os modelos de regressão aleatória e multi-característica foram semelhantes. Entretanto, sob o efeito da seleção, o modelo multi-característica mostrou-se mais susceptível ao viés de seleção.

Palavras-chave: Modelo de regressão aleatória, modelo multi-característica, modelos de repetibilidade, seleção.

REFERENCES

- Dal Zotto R (2000) Compararison of different test-day models for genetic evaluation of Italian Brown dairy cattle. **Proceedings of the international workshop on genetic improvement of functional traits in cattle** 25: 95-98.
- Henderson Junior CR (1982) Analysis of covariance in the mixed model: higher level, nonhomogeneous, and random regressions. **Biometrics** 38: 623-640.
- Kirkpatrick M, Lofsvold D and Bulmer M (1990) Analysis of the inheritance, selection and evolution of growth trajectories. **Genetics** 24: 979-993.
- Meyer K (1998) DXMRR - A set programs to estimate covariance functions for longitudinal data by REML. In: **Proceedings of the World Congress of Genetics Applied to Livestock Production**, 6, University of New England.

- Rao CR (1973) **Linear statistical inference and its applications**. 2ed, John Wiley & Sons, New York, 552p.
- Resende MDV, Rezende GDSP and Fernandes JSC (2001) Regressão aleatória e funções de covariância na análises de medidas repetidas. **Revista de Matemática e Estatística** 19: 21-40.
- SAS Institute Inc (1990) **SAS/STAT user's guide, version 6**. 4ed, Carry, 943p.
- Schaeffer LR and Dekkers JCM (1994) Random regression in animal models for test day production in dairy cattle. **Proceedings of the world congress genetic applied livestock production** 5: 443-446.
- Schaeffer LR and Wilton JW (1998) Comparison of single and multiple trait beef sire evaluation. **Canadian Journal of Animal Science** 76: 2303-2307.