Crop Breeding and Applied Biotechnology 8: 219-224, 2008
Brazilian Society of Plant Breeding. Printed in Brazil

Estimation and prediction of parameters and breeding values in soybean using REML/BLUP and Least Squares

Agnaldo Donizete Ferreira de Carvalho[1*], Roberto Fritsche Neto[1], and Isaias Olívio Geraldi[1]

Received 07 April 2007

Accepted 21 June 2008

ABSTRACT – *The aim of this study was to compare REML/BLUP and Least Square procedures in the prediction and estimation of genetic parameters and breeding values in soybean progenies. $F_{2:3}$ and $F_{4:5}$ progenies were evaluated in the 2005/06 growing season and the $F_{2:4}$ and $F_{4:6}$ generations derived thereof were evaluated in 2006/07. These progenies were originated from two semi-early experimental lines that differ in grain yield. The experiments were conducted in a lattice design and plots consisted of a 2 m row, spaced 0.5 m apart. The trait grain yield per plot was evaluated. It was observed that early selection is more efficient for the discrimination of the best lines from the F_4 generation onwards. No practical differences were observed between the least square and REML/BLUP procedures in the case of the models and simplifications for REML/BLUP used here.*

Key words: mixed models, early selection, *Glycine max*.

INTRODUCTION

Early selection or screening is used in the improvement of autogamous species to assess the potential of progenies at early stages ($F_{2:}$, $F_{3:}$ or $F_{4:}$). Progenies with a low potential for the traits of interest are therefore eliminated and the efforts focus on the potentially best genotypes (Fehr 1987). This method is based on the premise that the performance of a progeny in early generations is a good predictor of the performance of the inbred lines derived thereof (Bernardo 2003).

The linear mixed models based on REML/BLUP procedure modified the estimates of components of variance and genetic parameters (Resende et al. 1996). Previously, by the least square method, the covariances were estimated and interpreted in terms of their mathematical expectation (fitting them to the expected values), resulting in the components of variance. For this purpose, some assumptions had to be made: additivity model, normal data distribution, independence and homogeneity of errors. Currently, the variance components and the variances of random effects in

[1]Universidade de São Paulo, Escola Superior de Agricultura Luiz de Queiroz, Departamento de Genética, C.P. 83, 13400-970, Piracicaba, SP, Brazil. *E-mail: carvalhoadf@hotmail.com

linear mixed models can be estimated directly in the data set, where it is not necessary to meet the assumptions above.

Henderson (1974) cites two restrictions to the least square application: i) inability to estimate the breeding values of non-observed individuals; ii) only certain linear combinations of parameters can be estimated. Moreover, the linear mixed model is a flexible instrument in the estimation and prediction of genetic parameters and values, because it can be applied to unbalanced data from different generations.

Panter and Allen (1995) compared statistical procedures in the evaluation and selection of the best crosses in soybean, under different selection intensities and imbalance levels in the data set. They obtained correlation coefficients between the observed progeny means and predicted values of 0.74 and 0.61, by BLUP and least squares, respectively, indicating a greater consistency in BLUP.

In studies conducted by Resende et al. (1996), Farias Neto and Resende (2001) and Resende et al. (2001) in other plant species, procedures to estimate the components of variance and predict breeding values were compared and similar results were found for the effects of genotype ranking for ordinary least squares, generalized least squares, best prediction, best linear prediction, and the best linear unbiased prediction in situations with balanced data and homogeneity of variance, but not in the estimation of genetic variance components and prediction of genetic values. These authors claim that in any situation of balanced or unbalanced data, the components of variance estimation by restricted maximum likelihood (REML) and the breeding value prediction by the best linear unbiased prediction (BLUP) are equal or superior to the other procedures, so the REML/BLUP is currently recommended for studies in quantitative genetics.

The aim of this study was to compare the REML/BLUP and Least Square procedures in the prediction and estimation of genetic parameters and breeding values of soybean.

MATERIAL AND METHODS

Plant material and trial

Two semi-early experimental lines that differ in grain yield were used (lines 24-14 and 24-38), from a soybean breeding program of the department of genetics of the Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo. The cross between these two lines resulted in the F_1 generation, which was selfed naturally in a greenhouse to generate the F_2 generation seeds. These in turn, were sown in a greenhouse and harvested individually for the 100 $F_{2:3}$ progenies. Similarly, the $F_{4:5}$ generation was obtained. From each F_2 plant a pod was randomly taken as well, which was bulk-harvested, resulting in the F_3 and then the F_4 generations.

The 100 progenies of each of the $F_{2:3}$ and $F_{4:5}$ generations were the basic material for evaluation in the 2005/06 growing season, which were grown at the experimental station of the department of genetics of the ESALQ/USP. The generations $F_{2:4}$ and $F_{4:6}$ derived thereof were evaluated in the 2006/07 growing season, at the experimental stations of the department of genetics and of Anhembi, Piracicaba/SP (lat 22°43' S; long 47°36' W, 543 m asl). The experimental design used was the lattice square, with a triple replication of the $F_{4:5}$ and a double replication of the other generations. The experimental plots consisted of one row of 2 m with 0.5 between-row spacing. The cultural treatments during the experiments followed the technical recommendations for soybean in the state of São Paulo, Brazil. The trait evaluated was grain yield per plot. None of the experiments showed any degree data imbalance.

Genetic-statistical analysis by least squares

For the statistical analyses by the least square procedure the GLM Procedure (PROC GLM) of the computer system Statistical Analysis System (SAS) version 8.1 were used.

The statistical model used for individual analysis is shown below, where all effects, except the mean (μ), were considered random:

$$Y_{ijk} = \mu + t_i + r_j + b_{k(j)} + e_{ijk}$$

where,

Y_{ijk} is the observed value

μ is the general mean

t_i is the progeny effect (i = 1, 2, ... 100)

r_j is the replication effect (j = 1, 2 , ... 3 or 4)

$b_{k(j)}$ is the block effect within replications (k = 1, 2, ... 10)

e_{ijk} is the experimental error

The estimates of responses to selection (Rs) were obtained by the following estimator: $Rs = i\hat{h}^2\hat{\sigma}_F$ where

i is the selection index, based on standard phenotypic deviation; $\hat{h}^2$ is the broad-sense heritability and $\hat{\sigma}_{\bar{F}}$ is the standard phenotypic deviation.

The genetic correlations between generations were calculated according to the procedure described by Cruz and Regazzi (2001).

Genetic-statistical analysis by REML/BLUP

The fixed effects were estimated and the random effects predicted using the Mixed Procedure (PROC MIXED) of the computer system Statistical Analysis System (SAS) version 8.1, following a linear mixed model, described by Henderson (1984):

$y = Xr + Za + Wb + e$

where,

y is the phenotypic data vector

r is the vector of replication effects (fixed), added to the general mean

a is the vector of genetic effects (random), where, $a \sim N(0,G)$ and $G = A\sigma_a^2$

b is the vector of block effects (random)

e is the vector of residues (random), where, $e \sim N(0,R)$ and $R = I\sigma_e^2$.

The capital letters represent the matrix incidences for these effects, formed by values 0 and 1, which associate the unknown r, a and b with data vector y, respectively.

Vector r contemplates all replications of all places, in other words, the effects of location and replication within locations.

In the mixed models focus, G refers to the matrix of genetic covariances between the progenies, designated $A\sigma_a^2$. For A the parentage coefficient was disregarded here, therefore the matrix G was designated $I\sigma_a^2$, where $A=I$. Thus, σ_a^2 is equivalent to the genetic variance between progenies.

If the estimates of variances of the random effects are known, the fixed effects can be estimated and the random effects predicted simultaneously by the mixed model equation (MME) given by:

$$\begin{bmatrix} \hat{r} \\ \hat{a} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} X'X & X'Z & X'W \\ Z'X & Z'Z + A^{-1}\sigma_e^2/\sigma_a^2 & Z'W \\ W'X & W'Z & W'W + I\sigma_e^2/\sigma_b^2 \end{bmatrix} = \begin{bmatrix} Xy \\ Z'y \\ Wy \end{bmatrix}$$

For the above solutions, the genetic and non-genetic components of variance were considered unknown, which is a practical reality, and were estimated by the restricted maximum likelihood method (REML). Once the REML is an iterative process, the numerical algorithm known as Expectation Maximization (with alternating steps of expectation and maximization) was used, which characterizes the algorithm as EM-REML. Thus, from arbitrary initial values to $\hat{\sigma}_a^2$ and $\hat{\sigma}_b^2$ (block variance), solutions for $\hat{r}$, $\hat{a}$ and $\hat{b}$ are obtained. These solutions are used to obtain new estimates of the components of variance and so on, until the convergence is reached.

The response to selection by REML/BLUP was predicted by the mean breeding or genetic values of the selected progenies. The estimates of the genetic coefficient correlations between generations as well as the broad-sense heritability were calculated by the method proposed by Resende (2002).

The confidence intervals for the components of variance and the broad-sense heritability coefficients were obtained using the method proposed by Barbin (1993) and the method developed by Fisher, as described by Sokal and Rohlf (1995), was used for the correlation coefficients.

RESULTS AND DISCUSSION

In a comparison of the genetic variances estimated by both statistical procedures, those obtained by REML/BLUP in the generations evaluated in only one environment ($F_{2:3}$ and $F_{4:5}$) were higher. An opposite effect was observed in generations evaluated in two environments ($F_{2:4}$ and $F_{4:6}$), but did not differ significantly ($\alpha = 0.05$) when their confidence intervals were considered (Table 1).

The performance of the broad-sense heritability coefficients of the plot means and the estimates of genetic variances were the same by both statistical procedures. When the confidence intervals were considered the differences were not significant either ($\alpha = 0.05$) (Table 1). An important function of heritability, according to Falconer and Mackay (1996) is the role in genetic gain as estimator of the breeding value according to the selection expressed by the phenotypic value. However, better responses to selection are not necessarily associated with traits of high heritability.

Table 1. Estimates of genetic variance ($\hat{\sigma}_g^2$) and the broad-sense heritability ($\hat{h}^2$) obtained by the least squares and REML/BLUP procedures for the grain yield of progenies in different soybean generations

Generation	$\hat{\sigma}_g^2$		$\hat{h}^2$	
	REML/BLUP	Least Squares	REML/BLUP	Least Squares
$F_{2:3}$	322.32	277.50	0.37	0.33
	[254.32–443.95]*	[212.04–370.14]	[0.25–0.49]	[0.21–0.44]
$F_{2:4}$	275.92	348.05	0.42	0.53
	[210.84–368.04]	[265.95–464.24]	[0.33–0.52]	[0.42–0.63]
$F_{4:5}$	415.64	378.85	0.31	0.28
	[317.60–554.39]	[289.49–505.32]	[0.20–0.42]	[0.17–0.38]
$F_{4:6}$	465.50	485.28	0.52	0.62
	[355.70–620.90]	[370.81–647.28]	[0.51–0.74]	[0.58–0.81]

* Confidence intervals ($\alpha = 0.05$).

For example, the narrow-sense heritability can be high regardless of a low additive genetic variance, provided that the environmental effect is also small. For the evaluation of heritability as indicator of prediction, the breeder must above all know how much of the differential selection is expected in the gain. For traits of high heritability along with a high differential of selection, larger gains are therefore expected.

The estimates of genetic correlations (r_g) and their confidence intervals in the $F_{2:}$ and $F_{4:}$ generations and the generations derived thereof were highly significant ($p = 0.01$) and positive by the two statistical procedures (Table 2). In the literature, results of studies involving early generations to predict the performance of homozygous generations are inconsistent, and in many cases the genetic correlations between these estimates are low and insignificant (McKenzie and Lambert 1961, Briggs and Shebesky 1971, Knott and Kumar 1975). Probably the lack of r_g in early generations in early selection stages, such as $F_{2:}$ and the lines derived thereof is due, among other factors, to the high frequency of loci in heterozygosis and the high variability within progenies. Therefore, the testing and selection of progenies in more advanced generations such as F_4 seems more promising.

Table 2. Estimates of genetic correlations (r_g) obtained by least squares and REML/BLUP procedures for the grain yield of progenies in different soybean generations

Generation	REML/BLUP	Least Squares
$F_{2:3}/F_{2:4}$	0.34**	0.24**
	[0.16–0.51][a]	[0.05–0.42]
$F_{4:5}/F_{4:6}$	0.36**	0.38**
	[0.18–0.52]	[0.20–0.54]

** Significant at 1% probability by the t test. [a] Intervals of confidence (0.05).

Considering the estimated gain selection obtained by the two procedures and the gains based on the performance observed in advanced generations (Table 3), it was stated that under higher selection intensity in the $F_{2:3}$ generation, the REML/BLUP and least squares estimated similar gains for the $F_{2:4}$ generation. These values are however discrepant in relation to the observed selection gains. This discrepancy between the estimated and observed in both statistical procedures decreases as the selection index is reduced. At a selection intensity of over 30%, both estimate similar values to those observed. On the other hand, in the $F_{4:5}$ (selection) and $F_{4:6}$ generations (gain), the estimated results of both procedures were similar to each other and to the gains, at selection intensities between 10 and 40%. This shows that less biased estimates are obtained from the $F_{4:}$ generation. Moreover, this indicates that the response to selection in the $F_{2:}$ generation, in spite of efficient, was less accurate than in $F_{4:}$.

The response to selection depends on the estimates of genetic and phenotypic variances of a population. This has a considerable effect on early generations, since it is impossible to evaluate appropriate number of replicates and locations in breeding programs to ensure reliable estimates of the phenotypic and genetic parameters in these generations. It is further known that the two procedures (REML/BLUP and least squares) can result in rather divergent classifications when a set of treatments from different populations is considered, since the mixed model approach uses information on the genetic variability, even in the presence of orthogonality and

balancing (Duarte and Vencovsky 2001). Moreover, in normal situations of incomplete blocks, subject to planned or not planned imbalance (as is the case here, where the lattice design was used) even changes in classification within a same population are expected, which has a strong impact on selection.

Based on the results of our study, it was concluded that the two statistical procedures did not differ significantly ($\alpha = 0.05$) in any of the situations analyzed, which reinforces results of Farias Neto and Resende (2001). According to these authors, in the case of a balanced data set the REML/BLUP and least square procedures lead to identical results, whereas in cases of a slight imbalance results tend to be similar. However, further studies should be conducted in this direction, because the use of predictive techniques not only relevant, but also a viable possibility for time and cost savings in the evaluation of progenies in a breeding program.

CONCLUSIONS

No practical differences were observed between the least square and REML/BLUP procedures with the models and simplifications adopted for REML/BLUP here.

ACKNOWLEDGEMENTS

The authors thank the Genetics Department of ESALQ/USP, CAPES and CNPq.

Table 3. Estimates of responses to selection predicted and observed for the grain yield obtained by the least squares and REML/BLUP procedures

Response to selection		10%*	20%	30%	40%	50%
Predicted $F_{2:3}$	Least Squares	7.70	6.15	5.09	4.24	3.51
	REML/BLUP	7.65	6.72	5.85	4.98	4.19
Observed $F_{2:4}$		3.19	3.22	4.32	3.78	4.24
Predicted $F_{4:5}$	Least Squares	8.66	6.91	5.72	4.77	3.95
	REML/BLUP	9.36	7.44	6.08	5.17	4.37
Observed $F_{4:6}$		10.77	8.07	6.40	4.83	2.61

*Percentage of selected progenies.

Estimação e predição de parâmetros e valores genéticos em soja utilizando REML/BLUP e quadrados mínimos

RESUMO – *O objetivo deste trabalho foi comparar os procedimentos REML/BLUP e quadrados mínimos na estimação e predição de parâmetros e valores genéticos em progênies de soja. Foram avaliadas progênies $F_{2:3}$ e $F_{4:5}$ no ano agrícola 2005/06 e suas gerações derivadas $F_{2:4}$ e $F_{4:6}$ em 2006/07. Essas progênies são derivadas de duas linhagens experimentais de ciclo semiprecoce e contrastantes para a produção de grãos. Os experimentos foram delineados em látice, sendo a parcela experimental constituída de uma linha de 2 m espaçada de 0,5 m. O caráter avaliado foi a produção de grãos. Foi observado que o processo de seleção em gerações precoces apresenta maior eficiência na descriminação das melhores linhagens a partir da geração F_4. Sob os modelos e simplificações adotadas para o REML/BLUP, os procedimentos quadrados mínimos e REML/BLUP não apresentaram diferenças práticas.*

Palavras chave: modelos mistos, seleção precoce, *Glycine max*.

REFERENCES

Barbin D (1993) **Componentes de variância: teoria e aplicações**. Editora FEALQ, Piracicaba, 120p.

Bernardo R (2003) On the effectiveness of early generation selection in self-pollinated crops. **Crop Science 43**:1558-1560.

Briggs KG and Shebeski LH (1971) Early generation selection for yield and bread making quality of hard red spring wheat. **Euphytica 20**: 453-463.

Cruz CD and Regazzi AJ (2001) **Modelos Biométricos Aplicados ao Melhoramento Genético**. Editora UFV, Viçosa, 390p.

Duarte J and Vencovsky R (2001) Estimação e predição por modelo linear misto com ênfase na ordenação de médias de tratamentos genéticos. **Scientia Agrícola 58**: 109-117.

Falconer DS and Mackay TFC (1996) **Introduction to quantitative genetics.** Addison Wesley Longman Press, Harlow, 464p.

Farias Neto JT and Resende MDV (2001) Aplicação da metodologia de modelos mistos (REML/BLUP) na estimação de componentes de variância e predição de valores genéticos em pupunheira (*Bactris gasipaes*). **Revista Brasileira de Fruticultura 23**: 320-324.

Fehr WR (1987) **Principles of cultivar development**: Theory and Technique. Mc-Graw Hill Press, New York, 536 p

Henderson CR (1984) **Applications of linear models in animal breeding**. University of Guelph Press, Guelph, 462p.

Knott DR and Kumar J (1975) Comparison of early generation yield testing and a single seed descent procedure in wheat breeding. **Crop Science 15**: 295-299.

McKenzie RIH and Lambert JWA (1961) comparison of F_3 lines and their related F_6 lines in two barley crosses. **Crop Science 1**: 246-249.

Panter DM and Allen FL (1995) Using best linear unbiased predictions to enhance breeding for yield in soybean: II Selection of superior crosses from a limited number of yield trials. **Crop Science 35**: 405-410.

Resende MDV (2002) **Genética biométrica e estatística no melhoramento de plantas perenes**. Editora Embrapa Informação Tecnológica, Brasília, 975p.

Resende MDV, Furlani-Júnior E, Moraes MLT and Fazuoli LC (2001) Estimativas de parâmetros genéticos e predição de valores genotípicos no melhoramento do cafeeiro pelo procedimento REML/BLUP. **Bragantia 60**: 185-193.

Resende MDV, Prates DF, Jesus A and Yamada CK (1996) Estimação de componentes de variância e predição de valores genéticos pelo método da máxima verossimilhança restrita (REML) e melhor predição não viciada (BLUP) em *Pinus*. **Boletim de Pesquisa Florestal 32/33**: 23-42.

Sokal RR and Rohlf FJ (1995) **Biometry: the principles and practice of statistics in biological research.** (3rd). W. H. Freeman, New York, 887p.