Genetics and Molecular Biology, 30, 3, 536-544 (2007)

www.sbg.org.br

Research Article

Use of ridge regression for the prediction of early growth performance in crossbred calves

Eduardo da Cruz Gouveia Pimentel[1], Sandra Aidar de Queiroz[1], Roberto Carvalheiro[1,2] and Luiz Alberto Fries[2,3]

[1]*Departamento de Zootecnia, Faculdade de Ciências Agrárias e Veterinárias, Universidade Estadual Paulista, Jaboticabal, SP, Brazil.*
[2]*GenSys Consultores Associados S/S Ltda., Porto Alegre, RS, Brazil.*
[3]*Lagoa da Serra Ltda (Holland Genetics), Sertãozinho, SP, Brazil.*

Abstract

The problem of multicollinearity in regression analysis was studied. Ridge regression (RR) techniques were used to estimate parameters affecting the performance of crossbred calves raised in tropical and subtropical regions by a model including additive, dominance, joint additive or "profit heterosis" and epistatic effects and their interactions with latitude in an attempt to model genotype by environment interactions. A software was developed in Fortran 77 to perform five variant types of RR: the originally proposed method; the method implemented by SAS; and three methods of weighting the RR parameter λ. Three mathematical criteria were tested with the aim of choosing a value for the λ coefficient: the sum and the harmonic mean of the absolute Student t-values and the value of λ at which all variance inflation factors (VIF) became lower than 300. Prediction surfaces obtained from estimated coefficients were used to compare the five methods and three criteria. It was concluded that RR could be a good alternative to overcome multicollinearity problems. For all the methods tested, acceptable prediction surfaces could be obtained when the VIF criterion was employed. This mathematical criterion is thus recommended as an auxiliary tool for choosing λ.

Key words: crossbreeding, epistasis, genotype by environment interaction, heterosis, multicollinearity.

Received: February 23, 2006; Accepted: December 20, 2006.

Introduction

Many applications in animal breeding involve the prediction of one variable as a function of several others. The statistical technique most commonly employed for deriving prediction equations is ordinary least-squares regression analysis. When some of the explaining variables are highly correlated the ordinary least-squares predictor, albeit unbiased, may have large variances (Bergmann and Hohenboken, 1995). Ridge regression (RR) is an alternative technique to be employed when such ill-conditioning problems occur (Hoerl, 1962).

Standard models presently used for the genetic evaluation of crossbred beef cattle contain functions of up to four covariates to account for breed differences and heterosis: (direct and maternal) additive and dominance effects. Many recent studies have indicated that accounting for other genotypic effects may improve the accuracy of predicting performance of untested genotypes (Hirooka *et al.*, 1998; Arthur *et al.*, 1999; Kahi *et al.*, 2000; Mohamed *et al.*, 2001; Demeke *et al.*, 2003). According to Brito *et al.* (2002) and Piccoli *et al.* (2002) the following parameters should be considered in a proper model to design crossbreeding programs for specific regions, based on (pre) weaning traits: additive direct and maternal contributions from each breed; interactions between breeds in their additive contributions; direct and maternal dominance effects; direct and maternal epistatic effects; and interactions between these genotypic components and environmental variables.

In the estimation of crossbreeding parameters, when other genotypic effects (*e.g.* epistasis) are added to the additive-dominance model multicollinearity may be a problem as reported by Kinghorn and Vercoe (1989), Cassady *et al.* (2002) and Roso *et al.* (2005a).

The aim of this study was to compare different forms of implementation of RR to overcome effects of multicollinearity on the prediction of early growth performance in crossbred animals raised in different latitudes, us-

Send correspondence to Eduardo da Cruz Gouveia Pimentel. Departamento de Zootecnia, Faculdade de Ciências Agrárias e Veterinárias, Universidade Estadual Paulista, Via de Acesso Prof. Paulo Donato Castellane s/n, 14884-900 Jaboticabal, SP, Brazil. E-mail: pimentel@fcav.unesp.br.

ing the model proposed by Brito *et al.* (2002) and Piccoli *et al.* (2002).

Statistical background

Multicollinearity is defined as the existence of nearly linear dependency among columns of the design matrix **X** in the linear model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$. It induces numerical instability into the estimates and has dire consequences on their precision. It limits the size of the coefficient of determination and makes it increasingly more difficult to add unique explanatory prediction from additional variables. It also makes determining the contribution of each explanatory variable difficult because the effects of these variables are "mixed" or confounded due to collinearity. Coefficients may have the wrong sign or an implausible magnitude (Hair Jr *et al.*, 1992).

Several criteria have been used to detect multicollinearity problems. Draper and Smith (1998) suggest the following: (1) check if some regression coefficients have the wrong sign, based on prior knowledge; (2) check if predictors anticipated to be important based on prior knowledge have regression coefficients with small t-statistics; (3) check if deletion of a row or a column of the X matrix produces surprisingly large changes in the fitted model; (4) check the correlations between all pairs of predictor variables to see if any are surprisingly high; and (5) examine the *variance inflation factors* (VIF).

Let R_i^2 be the squared multiple correlation coefficient that result from the regression of x_i against all other explanatory variables. The *variance inflation* of x_i is then given by:

$$VIF_i = \frac{1}{1-R_i^2}.$$

It is clear that if x_i has a strong linear relation with other explanatory variables, R_i^2 is close to 1 and VIF values will tend to be very high. In the absence of any linear relation among explanatory variables, R_i^2 is zero and VIF equals 1. A VIF value greater than 1 indicates deviation from orthogonality and tendency to collinearity (Chatterjee and Price, 1991). There is no well-defined critical value that characterizes large VIF. Leclerc and Pireaux (1995) suggest that VIF values exceeding 300 may indicate the presence of possibly troublesome multicollinearity. Previous work by these authors (Pimentel *et al.*, 2004) has shown that this suggestion is valuable.

Examining a pairwise correlation matrix of explanatory variables might be insufficient to identify collinearity problems because near linear dependencies may exist among more complex combinations of regressors, that is, pairwise independence does not imply independence at all. Because VIF is a function of the multiple correlation coefficient among the explanatory variables, it is a much more informative tool for detecting multicollinearity than the simple pairwise correlations.

Many procedures have been suggested in an attempt to overcome the effects of multicollinearity in regression analysis. Freund and Wilson (1998) summarize them into three classes: variable selection, variable redefinition and biased estimation. These first two approaches actually represent some form of model reparametrization strategy for dealing with the problems caused by multicollinearity. Another alternative to be considered is the use of *a priori* information.

Lipovetsky and Conklin (2001) argue that, even in the presence of multicollinearity, it can be desirable to keep all possible variables in the model and to estimate their comparative importance in their relation to the response variable. This analysis strategy is justified because all available variables are not exact representation of each other. Rather each of the explanatory variables plays its own specific role in fitting and describing the behavior of the response variable.

Gau and Kohlhepp (1978) state that it should be recognized that any attempt to heuristically design simpler models with possibly lower levels of multicollinearity would explicitly require selection of certain variables which may lead to: (1) model specification errors; and (2) sacrificing of information. Every important variable, from a theoretical perspective or *a priori* knowledge, must be considered in model definition. When multicollinearity is a problem, biased estimation methods, such as RR, should be faced as the next step in the analysis procedure, after modeling measures, which include variable selection and/or redefinition. In the words of Leclerc and Pireaux (1995): *multicollinearity is not an inferential ground on which one can objectively reject a regression, a model or an estimate. The 'true' model may well be ill-conditioned after all.*

Ridge regression is characterized by the addition of small positive quantities (λ) to the diagonal elements of the coefficient matrix, as a means to reduce linear dependencies observed among its columns. A solution vector is thus obtained by the expression $\hat{\beta}^* = (X'X+\lambda I)^{-1} X'y$ where I is an identity matrix of the same order as X'X. Perhaps the best way for choosing the RR parameter (λ) would be the minimization of the expected squared difference between the estimate and the parameter being estimated (MSE). This would reveal the ideal balance between increase in bias and reduction in variance of the estimator (PEV), remembering that $MSE = PEV + bias^2$. Since bias estimation is left to simulation studies, analyses of real data demand other, sometimes subjective, arguments.

Some alternative forms of implementation of RR have recently been suggested. Aldrin (1997) present the *length modified RR*, which seems like a combination of James-Stein and RR: James-Stein estimators modify the ordinary least squares solution by shrinking the length of $\hat{\beta}$. Foucart (1999) present a *partial RR*, showing that the addition of a constant to some of the diagonal elements of the

matrix is sufficient for obtaining satisfactory estimates of the regression coefficients. The interest of such an approach, in contrast to classical RR, is that: by modifying only the diagonal terms involved in collinearity, one can get satisfying estimates of the regression coefficients overestimated by least-squares estimator, without perturbing the other regression coefficients.

Materials and Methods

Data and model

Data on 109,614 records of Hereford *(Bos taurus)* x Nelore (*Bos indicus*) calves born from 1974 to 1998, distributed in 4,665 contemporary groups (CG), and raised in 29 farms in the Brazilian states of Mato Grosso, Goiás, Mato Grosso do Sul, São Paulo, Paraná and Rio Grande do Sul, located between latitudes 14° S and 31.5° S, were used. Each of these farms also runs its own breeding program and they produce Hereford and Braford genetics (southern states) or Nelore and Braford genetics (central regions). This data set does not follow the general pattern of using imported and improved genetics on local unimproved populations.

A large proportion of calves (43.7%) were produced by artificial insemination, mostly from highly selected native Nelore and Hereford sires. Some Braford semen was imported, mainly from Argentina and most of these sires were 3/8 Brahman (*Bos indicus*). Given that Nelore took part in the formation of Brahman no distinction was made with respect to the origin of the Braford sires. The average sire had a Nelore composition of 0.21 ± 0.33.

Table 1 presents joint and marginal frequencies of records across latitudes and breed compositions of calves. Breed compositions were grouped in classes of eighths form Hereford (0) to Nelore (1).

The model related pre-weaning average daily gain (ADG) to direct and maternal additive (**da** and **ma**), dominance (**dd** and **md**), epistatic (**de** and **me**) and joint additive (**dc** and **mc**) effects. Covariates for **da** and **ma** were defined by expected Nelore contribution to the genetic make up of each calf and its dam. Dominance effects were estimated via direct and maternal coefficients of heterozygosity. Epistatic effects were calculated as the average heterozygosity present in the gametes which generated each offspring, that is, the average of the heterozygosity coefficients of the parents (**de**) and maternal grandparents (**me**) of an offspring (Fries *et al.*, 2000a; Roso *et al.*, 2005b). Coefficients of heterozygosity are relative to the maximum of 1.00 in generation F1, and coefficients of epistatic effects, as calculated here, are relative to the maximum of 1.00 in generation F2. Covariates for **dc** and **mc** were calculated as **dc** = **da***(1 - **da**) and **mc** = **ma***(1 - **ma**), respectively. When ancestral breed composition was not available, *inter se* mating was assumed. Sires, dams, maternal grandsires and maternal granddams had complete information to calculate their heterozygosities with the following frequencies: 0.76, 0.76, 0.74, and 0.73, respectively.

Contemporary groups were defined as farm x year of birth x sex x management group and julian weaning date. Attempting to save processing time and computational costs, a previous absorption (Searle, 1971) of the CG effect was carried out. Previous absorption of CG effects had removed the main effects of latitude, since animals in the same CG were all raised at the same latitude. Consequently, when a covariate for latitude was included in the model as a main effect, null parameter estimates were obtained. Latitude was then included in the model as a "modifier agent" in the form of interactions between genotypic and linear and quadratic latitude effects.

After absorption of CG effect, the final regression model used to analyze the data was as follows:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \mathbf{W}\delta + \varepsilon,$$

with

$$\mathbf{E}(\varepsilon) = \mathbf{0} \text{ and } V(\mathbf{y}) = V(\varepsilon) = \sigma^2 I_n,$$

Table 1 - Frequencies of records across latitudes and breed compositions of the calves, grouped in classes of eighths and expressed as the expected proportion of Nelore.

Genotype range	Latitude (° S)										Totals
	14.0	16.0	21.0	21.5	22.0	23.0	29.0	30.0	31.0	31.5	
0.000 to 0.124	15	5	-	2,287	2	2,202	30	13,870	31,114	1,783	51,308
0.125 to 0.249	-	-	-	-	2	-	72	147	370	67	658
0.250 to 0.374	9	36	9	-	90	336	723	3,319	3,595	1,684	9,801
0.375 to 0.499	37	15	54	-	155	-	983	6,284	4,662	2,704	14,894
0.500 to 0.624	2,195	12,383	1,359	-	2	695	180	1,162	1,817	634	20,427
0.625 to 0.749	1,845	46	163	-	-	-	347	65	58	186	2,710
0.750 to 0.874	1,541	989	322	-	-	11	331	12	307	65	3,578
0.875 to 1.000	449	5,320	312	-	-	49	71	3	3	31	6,238
Totals	6,091	18,794	2,219	2,287	251	3,293	2,737	24,862	41,926	7,154	109,614

where **y** is the vector of observed ADG; β is a vector of unknowns corresponding to the other environmental effects included in the model as covariables: linear and quadratic effects of age of dam nested in sex of calf, linear and quadratic effects of age of calf, and quadratic-quadratic spline function of julian date of birth; γ is the vector of unknowns corresponding to the eight genotypic effects considered; δ is the vector of unknowns for the interactions between the eight genotypic effects and linear and quadratic latitude effects; **X**, **Z** and **W** are the respective data matrices; and ε is a stochastic disturbance term.

Table 2 presents the basic descriptive statistics for the response and explanatory variables. By definition, joint additive coefficients range from 0.00 for purebreds to 0.25 for F1 offspring. All other genotypic covariates varied from 0.00 to 1.00, with the exception of **me**, for which the maximum value was 0.75. This means the data set contained F2 offspring but no F2 dams.

Statistical procedures

Multicollinearity diagnosis was performed using VIF values, eigenvalues of the correlation matrix and the condition number. The VIF values for each covariate were calculated as the product of diagonal element of the coefficient matrix, with respect to the covariate in question, by the corresponding diagonal element of the inverse of the coefficient matrix, as in Maindonald (1984). Eigenvalues of the correlation matrix and the condition number were obtained using the COLLINOINT option of the REG procedure of the SAS program (SAS Inst., Inc., Cary, NC).

A Fortran 77 program was developed to perform some variants of ridged multiple linear regression analysis and test three criteria for choosing the RR parameter (λ). Five implementation methods of RR were performed:

$$\hat{\beta}_{RR} = (X'X+\lambda I)^{-1} X'y. \quad (1)$$

The original implementation of RR, as presented by Hoerl (1962).

$$\hat{\beta}_{SAS} = \left[R_{XX} (I+D_{\lambda})\right]^{-1} R_{Xy}. \quad (2)$$

The RR technique implemented by SAS, as described by Freund and Littell (2000), where R_{XX} is the correlation matrix of explanatory variables, R_{Xy} is the vector of correlations between the response and the explanatory variables, and D_{λ} is a diagonal matrix with λ as the value of the diagonal elements.

$$\hat{\beta}_{WRR} = (X'X+\lambda W)^{-1} X'y, \quad (3)$$

where W is a diagonal matrix whose elements are the quotients of the diagonal elements of X'X by the smallest element of it. In this method, λ values are weighted by the ratio of the sum of squares of each covariate by the smallest sum of squares, in an attempt to adjust for the magnitude of each element of X'X.

$$\hat{\beta}_{VIF} = (X'X+\lambda F)^{-1} X'y, \quad (4)$$

where F is a diagonal matrix, such that $F = \{VIF_i\}$ if VIF_i is greater than 300 and $F = \{0\}$ otherwise. Here, the biasing factor (λ) is adjusted for the degree of involvement of each covariate in multicollinearity. The VIF values employed on weighting the RR parameter were the ones found for each covariate at $\lambda = 0$.

Table 2 - Means, standard deviations (SD), minimum and maximum values for the co-variables included in the record of each calf. The total number of observations was 109,614.

Variable [1]	Mean	SD	Minimum	Maximum
Average daily gain (kg/day)	0.67	0.16	0.25	1.50
Age of dam (years)	5.5	2.6	2.0	20.0
Weaning age (days)	201.2	29.6	100.0	320.0
Julian date of birth	269.0	49.1	1.0	365.0
Direct additive (**da**)	0.27	0.29	0.00	1.00
Maternal additive (**ma**)	0.34	0.40	0.00	1.00
Direct joint additive (**dc**)	0.11	0.12	0.00	0.25
Maternal joint additive (**mc**)	0.07	0.11	0.00	0.25
Direct dominance (**dd**)	0.31	0.36	0.00	1.00
Maternal dominance (**md**)	0.15	0.24	0.00	1.00
Direct epistasis (**de**)	0.13	0.19	0.00	1.00
Maternal epistasis (**me**)	0.12	0.20	0.00	0.75
Latitude (° S)	26.6	6.4	14.0	31.0

[1]**da** and **ma** = expected proportion of Nelore in the calf and its dam; **dc** and **mc** = products [**da***(1-**da**)] and [**ma***(1-**ma**)]; **dd** and **md** = expected breed heterozygosity (relative to the maximum of 1.00 in the F1) in calf and dam; **de** and **me** = average of estimated heterozygosities in parents and maternal grandparents of the calf.

$$\hat{\beta}_{CFW} = (X'X+\lambda C)^{-1} X'y, \quad (5)$$

where C is a diagonal matrix whose elements are the products of the corresponding elements of W and F. This combination of methods 3 and 4 is intended to adjust λ values for the magnitude of the elements of X'X and its involvement in multicollinearity, simultaneously.

The program was developed in a way to perform RR by the inclusion of a matrix of pseudo-observations (Lawson and Hanson, 1995) in each analysis, and for each range of λ values to be tested. If there are *n* observations, represented by $y_{n\times1}$, and *p* explanatory variables ($X_{n\times p}$), then the original data matrix is:

$$\begin{bmatrix} X_{n\times p} & y_{n\times1} \\ 0 & 0 \end{bmatrix}.$$

Then, given that S is a diagonal matrix whose diagonal entries equal the square root of λ (for λ being a given real number), it is necessary to add *p* pseudo-observations to form the new data set:

$$\begin{bmatrix} X_{n\times p} & y_{n\times1} \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ S_{p\times p} & 0_{p\times1} \end{bmatrix} = \begin{bmatrix} X_{n\times p} & y_{n\times1} \\ S_{p\times p} & 0_{p\times1} \end{bmatrix}.$$

Direct solutions to multiple regression problems will begin by triangularizing the following matrix:

$$\begin{bmatrix} X' & S' \\ y' & 0' \end{bmatrix} \times \begin{bmatrix} X & y \\ S & 0 \end{bmatrix} = \begin{bmatrix} X'X+S'S & X'y+S'0 \\ y'X+0'S & y'y+0'0 \end{bmatrix} = \begin{bmatrix} X'X+S'S & X'y \\ y'X & y'y \end{bmatrix}.$$

The coefficient matrices of real observations (X'X) were all the same, and thus all the variables contemplated in the model were defined in the same way, in all methods, except in method 2. What was actually changing from one method to another was the matrix of pseudo-observations that was added to the data set. In method 1 we had S'S = λI; in method 3, S'S = λW; and so forth. Following Marquardt (1970), the RR estimator is then seen to be a type of weighted average between the actual data and other data (in Bayesian terms, the "prior information") for which the response values are arbitrarily set to zero.

Method 2 operates on standardized variables and the change in notation (R_{XX} and R_{Xy}) was made in an attempt to characterize that they are correlation matrices. D_λ is actually the matrix λI. It must be emphasized that the vector of estimates in this method is defined by a different equation, which makes evident that off-diagonal entries are added to the correlation matrix.

Three mathematical criteria for choosing λ were evaluated: the sum and the harmonic mean of absolute Student-t values of the variables involved in collinearity, and their VIF. A maximum value for the sum of the absolute t-values (**sat**) may indicate the maximum reduction in standard errors of the estimates, with the possibly lowest λ (Fries *et al.*, 2000b). A maximum value for the harmonic mean of the absolute t-values (**hat**) may suggest higher uniformity in terms of statistically significant estimates (Piccoli *et al.*, 2002). The VIF criterion consisted of the choice of λ at which all VIF values were lower than 300 (Leclerc and Pireaux 1995).

Comparisons of the five types of implementation and three criteria were made under the exam of the prediction surfaces described by the estimates obtained for each implementation and criterion. Predicted values were calculated for nine genotypes, from pure Hereford to pure Nelore by eighths, at six latitudes: 16, 19, 22, 25, 28, and 31° S. All predicted values were added by a constant (0.61 kg/day) corresponding to an average environmental effect.

Results

The presence of multicollinearity was clearly evident in this analysis. Nine out of the 33 eigenvalues (9 environmental + 8 genotypic + 16 interaction covariates) were lower than 0.001 showing how much the correlation matrix approached singularity. The condition number, which can be defined by the ratio of the largest to the smallest eigenvalue, is used here as another reference for revealing this situation. Belsley *et al.* (1980) suggest that moderate to strong relations are associated with condition numbers of 30 to 100. The one observed in this analysis was 2969, which is almost a hundred times what is considered to be critical, having in mind that COLLINOINT option outputs the square root of this ratio (see Freund and Littell, 2000). Some VIF values were in the hundreds of thousands, characterizing intense association among the explanatory variables.

Figure 1 shows the primary ("subjective") criterion for choosing the RR parameter λ, which is done by plotting the values of estimated coefficients, referred to as the ridge trace (a), and respective Student t-values (b) against successive values of λ, for implementation methods 1 to 5. Note that, for better visualization, the range of λ values varies from one method to another.

Figure 2 shows the prediction surfaces described by estimated coefficients obtained when **sat** (a), **hat** (b) and VIF (c) criterion was used for choosing λ, for each of the five RR methods. Genotypes range from Hereford (0) to Nelore (1).

Discussion

Figure 1b1 shows an increase in absolute t-values can be observed until a certain point, after which they begin to decrease. The largest values for **sat** and **hat** were obtained at λ = 0.08 and λ = 0.06, respectively. For all breed compositions the predicted ADG values were lower at intermedi-

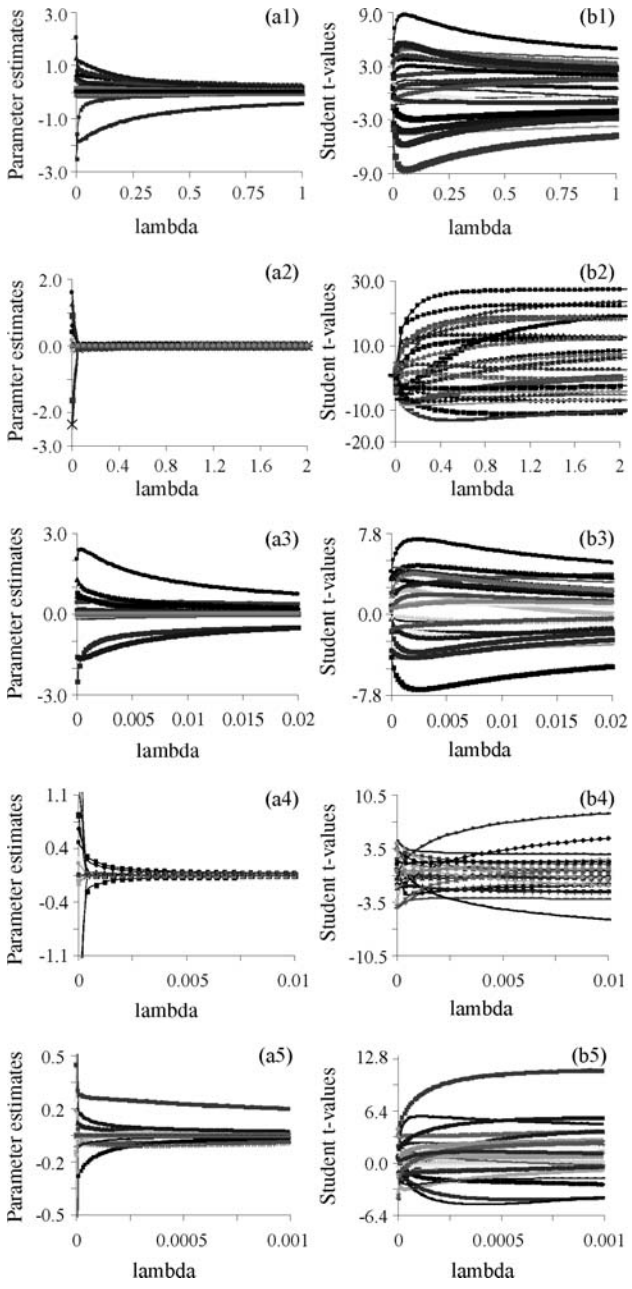

Figure 1 - Parameter estimates (a) and respective Student t-values (b) against successive values of λ, for the implementation of methods 1 to 5.

ate latitudes than at extreme latitudes (Figures 2a1 and 2b1). This is particularly evident for the Nelore genotype. It is very unlike that *B. indicus* cattle, selected to produce in tropical environments, would show greater performance in the state of Rio Grande do Sul (latitude 31° S – subtropical Brazilian region) than in São Paulo (latitude 23° S).

Looking at the ridge trace in Figure 1a1 we can see that the changes in the estimated coefficients still continue as λ increases towards values much larger than 0.06 or 0.08. It may suggest that using **sat** or **hat** criteria would not provide a λ value that represents enough change in estimated coefficients. In this case, not enough to provide prediction surfaces that meet the expectations from a biological perspective. In fact, the prediction surfaces obtained with 0.06 and 0.08 (Figures 2a1 and 2b1) are the same as the one for λ = 0, that is, for ordinary least-squares regression. One could speculate that the lack of practical coherence shown by these surfaces could be one of the reasons why attempts to account for other than additive-dominance genotypic effects are usually discarded.

An effective reduction in VIF occurs at much greater λ values. It is only at λ = 190 that they all become lower than 300 (Table 3). The prediction surface described for this λ (Figure 2c1) shows a continuous decrease in the predicted ADG of Nelore calves, as latitude gets higher. Better performance of crossbred animals compared to purebreds confirms a beneficial heterotic effect at all latitudes except at lower ones, where Nelore calves did better. Barlow (1981) observed that heterosis for growth among ruminants appears to be favored by a benign nutritional environment. Under low input conditions, genes for adaptation seem to have much more importance than genes for high metabolism, aggregated to genotypes by Hereford proportion, plus heterotic effects.

Method 2 provided very different results from method 1. Figures 1a2 and 1b2 show a sharp decrease in parameter estimates at very small λ values, while t-values keep rising. Largest **sat** and **hat** occur at λ = 2.65 and λ = 0.35, respectively. A drastic reduction in VIF is verified. They are all lower than 300 at λ = 0.0004 (Table 3).

This greater impact of RR implemented by SAS on parameter estimates can be explained by the fact that, in this method, elements are added to the whole correlation matrix, not only to its diagonal entries. In this way, it may possibly introduce a larger amount of bias in the analysis. Another point is related to variable standardization. Since SAS performs the addition of RR parameters on correlation matrices, smaller λ values may produce a greater effect than what could be expected on coefficient matrices. One may also have in mind that, in this method, λ values are added to elements corresponding to all covariates included in the model, with no distinction about their involvement in multicollinearity.

Figures 2a2 and 2b2 show that, when **sat** and **hat** criteria were used, this method of RR flattened the prediction surfaces. These criteria may have provided too large λ values, resulting in an excessive shrinkage of parameter estimates.

The surface corresponding to the VIF criterion (Figure 2c2) is very similar to the one obtained for the same criterion in method 1, except that it shows a slightly larger advantage for Nelore calves at lower latitudes. Since the greater impact of this method over parameter estimates is also valid for VIF values, the criterion based on them led to a more acceptable surface, and its shape corresponded to a more realistic situation.

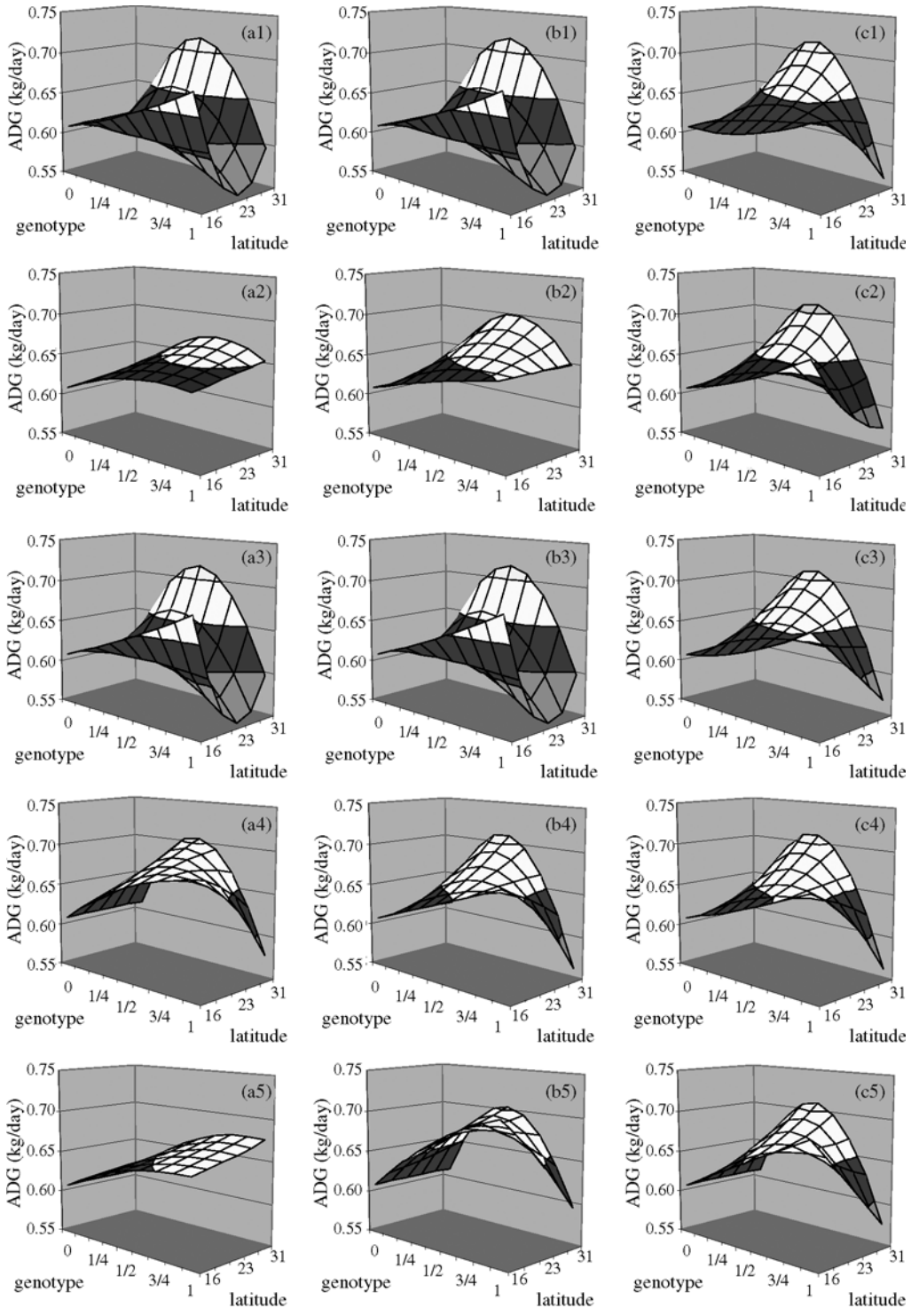

Figure 2 - Predicted average daily gain (ADG) surfaces described by estimated coefficients obtained according to the sum of absolute t-values (**sat)** (a), the harmonic mean of the absolute t-values (**hat)** (b) and the variance inflation factor (VIF) (c) criterion, for implementation of methods 1 to 5. Genotypes ranged from Hereford (0) to Nelore (1).

In method 3, the RR parameter was weighted according to the size of each diagonal entry of X'X (a sort of standardization process). All genotypic covariates are expressed in percentage, ranging from 0 to 1, except for joint additive effects, for which maximum value is 0.25. This may be the reason why the results obtained with this method are almost the same as those from method 1 (compare first and third lines of Figures 1 and 2). Values of λ were 0.0022, 0.0014 and 0.26 according to **sat**, **hat** and VIF criterion, respectively. The effects of this kind of adjust-

Table 3 - Maximum variance inflation factor (VIF) values according to the ridge parameter (λ) for each implementation method of ridge regression.

Method 1		Method 2		Method 3		Method 4		Method 5	
λ	maxVIF	λ	maxVIF	λ	maxVIF	λ	maxVIF	λ	maxVIF
0.0	371,906.4	0.0000	371,906.4	0.00	371,906.4	0.0000	371,906.4	0.00000	371,906.4
46.0	1,090.1	0.0001	1,434.3	0.06	1,102.8	0.0024	714.3	0.00001	1,004.7
92.0	590.9	0.0002	680.8	0.12	585.9	0.0048	454.2	0.00002	559.5
138.0	405.6	0.0003	404.5	0.18	400.8	0.0072	354.9	0.00003	387.9
190.0	299.6	0.0004	269.6	0.26	282.9	0.0100	293.3	0.00004	296.9

ment could possibly be better evidenced with other data sets, where covariates are expressed in more divergent units of measurement.

Method 4 included an adjustment of λ values according to the involvement of each covariate in collinearity. Maximum **sat** and **hat** were reached at λ = 0.29 and λ = 0.016, respectively. All VIF became lower than 300 at λ = 0.01 (Table 3). It can be seen in Figures 2a4, 2b4 and 2c4 that, in this method, there could be observed an agreement amongst the three criteria, specially **hat** and VIF. These surfaces are very close to the ones achieved when the VIF criterion was employed in methods 1 to 3, making method 4 the most robust one with respect to the criteria for choosing λ.

The expectation (at least ours) about method 5, which is a combination of 3 and 4, was that it would repeat the same path observed in method 4, but for lower λ values. Instead, it occurred only for VIF and **hat** (to some extent). The λ values defined by **sat**, **hat** and VIF criterion were 0.0275, 0.00022 and 0.00004, respectively. Perhaps the huge magnitude of the weights employed in this method had "broken down" the factor that provided such a coincidence amongst criteria in method 4. Vinod (1976) was probably right when he stated that "finding a λ for any specific problem remains something of an art".

The results of this work indicate that there is always some value of λ that improve regression analysis output (estimated coefficients or predicted values). This value may coincide with the point where the relationship between error variance and bias reaches its best configuration, as suggested by Hoerl and Kennard (1970). Variation of the value used in the five methods actually comes from the difference in magnitude of weights employed in each of them. Except in method 1, what was added to the coefficient matrix was, in fact, larger than the λ value informed to the program.

As pointed out by Marquardt (1970), the RR and the generalized inverse estimators share many properties: both are superior to least-squares for ill-conditioned problems; for both classes of estimators the degree of bias can be bracketed within a reasonable range in any given instance and practical results can be obtained. The generalized inverse solution is especially relevant for precisely zero eigenvalues. The RR solution is computationally simpler and it seems better suited to coping with very small, but nonzero, eigenvalues.

In the analysis described in the present paper, with this particular data set the VIF criterion has emerged as the most robust way for choosing λ, although consideration of the classic reference (current estimates) is not discarded. Here, examination of the signs and values of the estimated coefficients was unfeasible because of the inclusion of interaction terms between latitude and genotypic effects.

In our study, the key point was to look at prediction surfaces and carefully check that they were reasonable. This procedure can be interpreted from a Bayesian point of view as the specification of 5 or 10 prior distributions covering a plausible range followed by the comparison of the closeness of various alternative regression procedures to the 5 or 10 posterior means. Both approaches offer some clarification of the real problem posed by a given data set: if the correct analysis depends critically on the model and the prior information adopted, over some reasonable range, the researcher should not expect any specific procedure to be automatically applicable (Dempster *et al.* 1977).

In conclusion, ridge regression (RR) can be a good alternative to overcome multicollinearity problems not only when the interest is interpretation of signs and values of estimates but also when regression analysis is made for prediction purposes. The fourth method of implementation of RR was the most robust one, with respect to the criterion for choosing the RR parameter λ. Nevertheless, when the VIF criterion was used, all methods provided prediction surfaces showing quite acceptable interpretation from a biological perspective. This mathematical criterion for choosing λ is thus recommended as an indicator tool and should not exclude an examination of the signs and values of the estimated coefficients and a good understanding of the phenomenon under study.

Acknowledgments

The authors thank Conexão Delta G for provision of the data and Gensys Consultores Associados S/s Ltda. for data preparation. This work was supported by the Brazilian agencies CAPES, CNPq and FAPESP.

References

Aldrin M (1997) Length modified ridge regression. Comput Statist Data Anal 25:377-398.

Arthur PF, Hearnshaw H and Stephenson PD (1999) Direct and maternal additive and heterosis effects from crossing *Bos indicus* and *Bos taurus* cattle: Cow and calf performance in two environments. Livest Prod Sci 57:231-241.

Barlow R (1981) Experimental evidence for interaction between heterosis and environment in animals. Anim Breed Abstr 49:715-737.

Belsley DA, Kuh E and Welsch RE (1980) Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. 1st edition. John Wiley & Sons, New York, 292 pp.

Bergmann JAG and Hohenboken WD (1995) Alternatives to least squares in multiple linear regression to predict production traits. J Anim Breed Genet 112:1-16.

Brito FV, Piccoli ML, Severo JLP, Schenkel FS, Roso VM and Fries LA (2002) Estimating environmental and genotypic effects on preweaning weight gain of Angus x Nelore calves. Proc 7th World Cong Genet Appl Livest Prod, Montpellier, France, CD-ROM.

Cassady JP, Young LD and Leymaster KA (2002) Heterosis and recombination effects on pig growth and carcass traits. J Anim Sci 80:2286-2302.

Chatterjee S and Price B (1991) Regression Analysis by Example. 2nd edition. John Wiley & Sons, New York, 278 pp.

Demeke S, Neser FWC and Schoeman SJ (2003) Early growth performance of *Bos taurus* x *Bos indicus* cattle crosses in Ethiopia: Evaluation of different crossbreeding models. J Anim Breed Genet 120:39-50.

Dempster AP, Schatzoff M and Wermuth N (1977) A simulation study of alternatives to ordinary least squares. J Amer Statist Assoc 72:77-91.

Draper NR and Smith H (1998) Applied Regression Analysis. 3rd edition. John Wiley & Sons, New York, 706 pp.

Foucart T (1999) Stability of the inverse correlation matrix. Partial ridge regression. J Statist Plann Inference 77:141-154.

Freund RJ and Littell RC (2000) SAS System for Regression. 3rd edition. SAS Institute, Cary, 235 pp.

Freund RJ and Wilson WJ (1998) Regression Analysis: Statistical Modeling of a Response Variable. 1st edition. Academic Press, San Diego, 444 pp.

Fries LA, Johnston DJ, Hearnshaw H and Graser HU (2000a) Evidence of epistatic effects on weaning weight in crossbred beef cattle. Asian-Aust J Anim Sci 13:242 (Abstract).

Fries LA, Graser HU, Johnston DJ and Hearnshaw H (2000b) Using ridge regression to estimate genetic effects in crossbred beef cattle. Asian-Aust J Anim Sci 13:241 (Abstract).

Gau GW and Kohlhepp DB (1978) Multicollinearity and reduced-form price equations for residential markets: An evaluation of alternative estimation methods. Am R Estate Urban Econ Assoc J 6:50-69.

Hair Jr JF, Anderson RE, Tatham RL and Black WC (1992) Multivariate Data Analysis. 3rd edition. Macmillan Publishing Company, New York, 544 pp.

Hirooka H, Groen AF and Van der Werf JHJ (1998) Estimation of additive and non-additive genetic parameters for carcass traits on bulls in dairy, dual purpose and beef cattle breeds. Livest Prod Sci 54:99-105.

Hoerl AE (1962) Application of ridge analysis to regression problems. Chem Engineer Prog 58:54-59.

Hoerl AE and Kennard RW (1970) Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12:55-67.

Kahi AK, Thorpe W, Nitter G and Baker RL (2000) Crossbreeding for dairy production in the lowland tropics of Kenya. I. Estimation of individual crossbreeding effects on milk production and reproductive traits and on cow live weight. Livest Prod Sci 63:39-54.

Kinghorn BP and Vercoe PE (1989) The effects of using the wrong genetic model to predict the merit of crossbred genotypes. Anim Prod 49:209-216.

Lawson CL and Hanson RJ (1974) Solving Least Squares Problems. 1st edition. Prentice-Hall, Englewood Cliffs, 340 pp.

Leclerc G and Pireaux JJ (1995) The use of least squares for XPS peak parameters estimation. Part 3. Multicollinearity, ill-conditioning and constraint-induced bias. J Electron Spectrosc Relat Phenom 71:179-190.

Lipovetsky S and Conklin WM (2001) Multiobjective regression modifications for collinearity. Comput Oper Res 28:1333-1345.

Maindonald JH (1984) Statistical Computation. 1st edition. John Wiley & Sons, New York, 370 pp.

Marquardt DW (1970) Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. Technometrics 12:591-612.

Mohamed SA, Rottmann O and Pirchner F (2001) Components of heterosis for growth traits and litter size in line crosses of mice after long-term selection. J Anim Breed Genet 118:263-270.

Piccoli ML, Roso VM, Brito FV, Severo JLP, Schenkel FS and Fries LA (2002) Additive, complementarity (additive*additive), dominance, and epistatic effects on preweaning weight gain of Hereford x Nelore calves. Proc 7th World Cong Genet Appl Livest Prod, Montpellier, France, CD-ROM.

Pimentel ECG, Queiroz SA, Carvalheiro R and Fries LA (2004) Including epistasis and complementarity in models for genetic effects evaluation in crossbred beef cattle. Proc 5th Natl Symp Brazilian Soc Anim Breed, Pirassununga, Brazil, CD-ROM.

Roso VM, Schenkel FS, Miller SP and Schaeffer LR (2005a) Estimation of genetic effects in the presence of multicollinearity in multibreed beef cattle evaluation. J Anim Sci 83:1788-1800.

Roso VM, Schenkel FS, Miller SP and Wilton JW (2005b) Additive, dominance, and epistatic loss effects on preweaning weight gain of crossbred beef cattle from different *Bos taurus* breeds. J Anim Sci 83:1780-1787.

Searle SR (1971) Linear Models. 1st edition. John Wiley & Sons, New York, 532 pp.

Vinod HD (1976) Application of new ridge regression methods to a study of bell system scale economies. J Amer Statist Assoc 71:835-841.

Associate Editor: Pedro Franklin Barbosa