Degree of multicollinearity and variables involved in linear dependence in additive-dominant models

Juliana Petrini[1], Raphael Antonio Prado Dias[2], Simone Fernanda Nedel Pertile[1], Joanir Pereira Eler[3], José Bento Sterman Ferraz[3] and Gerson Barreto Mourão[1]

[1]Universidade de São Paulo (USP), Escola Superior de Agricultura Luiz de Queiroz, Departamento de Zootecnia, Avenida Pádua Dias, nº 11, Caixa Postal 9, Bairro Agronomia, CEP 13418-900 Piracicaba, SP, Brazil. E-mail: juliana.petrini@usp.br, spertile@usp.br, gbmourao@usp.br [2]Instituto Federal de Educação, Ciência e Tecnologia – Sul de Minas Gerais, Estrada de Muzambinho, Km 35, Bairro Morro Preto, CEP 37890-000 Inconfidentes, MG, Brazil. E-mail: raphael.a.p.dias@gmail.com [3]USP, Faculdade de Zootecnia e Engenharia de Alimentos, Departamento de Ciências Básicas, Avenida Duque de Caxias Norte, nº 225, CEP 13635-900 Pirassununga, SP, Brazil. E-mail: joapeler@usp.br, jbferraz@usp.br

Abstract – The objective of this work was to assess the degree of multicollinearity and to identify the variables involved in linear dependence relations in additive-dominant models. Data of birth weight (n=141,567), yearling weight (n=58,124), and scrotal circumference (n=20,371) of Montana Tropical composite cattle were used. Diagnosis of multicollinearity was based on the variance inflation factor (VIF) and on the evaluation of the condition indexes and eigenvalues from the correlation matrix among explanatory variables. The first model studied (RM) included the fixed effect of dam age class at calving and the covariates associated to the direct and maternal additive and non-additive effects. The second model (R) included all the effects of the RM model except the maternal additive effects. Multicollinearity was detected in both models for all traits considered, with VIF values of 1.03–70.20 for RM and 1.03–60.70 for R. Collinearity increased with the increase of variables in the model and the decrease in the number of observations, and it was classified as weak, with condition index values between 10.00 and 26.77. In general, the variables associated with additive and non-additive effects were involved in multicollinearity, partially due to the natural connection between these covariables as fractions of the biological types in breed composition.

Index terms: *Bos taurus* x *Bos indicus*, animal breeding, beef cattle, correlation matrix, crossbreeding, variance inflation factor.

Grau de multicolinearidade e variáveis envolvidas na dependência linear em modelos aditivo-dominantes

Resumo – O objetivo deste trabalho foi avaliar o grau de multicolinearidade e identificar as variáveis envolvidas na dependência linear em modelos aditivo-dominantes. Foram utilizados dados de peso ao nascimento (n=141.567), peso ao ano (n=58.124) e perímetro escrotal (n=20.371) de bovinos de corte compostos Montana Tropical. O diagnóstico de multicolinearidade foi baseado no fator de inflação de variância (VIF) e no exame dos índices de condição e dos autovalores da matriz de correlações entre as variáveis explanatórias. O primeiro modelo estudado (RM) incluiu o efeito fixo de classe de idade da mãe ao parto e as covariáveis associadas aos efeitos aditivos e não aditivos diretos e maternos. O segundo modelo (R) incluiu todos os efeitos do RM, exceto os efeitos aditivos maternos. Detectou-se multicolinearidade em ambos os modelos para todas as características consideradas, com valores de VIF de 1,03–70,20, para RM, e de 1,03–60,70, para R. As colinearidades aumentaram com o aumento de variáveis no modelo e com a redução no volume de observações, e foram classificadas como fracas, com valores de índice de condição entre 10,00 e 26,77. Em geral, as variáveis associadas aos efeitos aditivos e não aditivos estiveram envolvidas em multicolinearidade, parcialmente em razão da ligação natural entre essas covariáveis como frações dos tipos biológicos na composição racial.

Termos para indexação: *Bos taurus* x *Bos indicus*, melhoramento animal, bovino de corte, matriz de correlação, cruzamento, fator de inflação da variância.

Introduction

In animal breeding studies, an obstacle to obtaining reliable results is the presence of linear correlations between explanatory variables, which is defined as multicollinearity. Multicollinearity is caused mainly by physical restrictions in the model or population, due to sampling techniques or to a model with excessive terms (Mason et al., 1975). In this situation, the ordinary least squares method – an important methodology

used to estimate genetic parameters – yields unstable regression coefficients with large standard errors, leading to erroneous inferences (Bergmann & Hohenboken, 1995). Collinearity also makes the model outputs sensitive to changes in the database and to the addition or reduction of variables in the model (Belsley, 1991). Moreover, it results in high variances, which are detrimental to the use of hypothesis tests for regression coefficients, estimation, and prediction (Mansfield & Helms, 1982).

Problems related to multicollinearity in models for estimation of genetic effects in crossbred populations were reported in several studies (Cassady et al., 2002; Roso et al., 2005a; Pimentel et al., 2006; Toral et al., 2009; Lopes et al., 2010). Rodríguez-Almeida et al. (1997), in a study about direct and maternal additive effects for birth and weaning weights in multiracial populations, identified the presence of multicollinearity between direct and maternal heterosis, and between direct and maternal additive effects of the same biological type. Similarly, Roso et al. (2005b), working with purebred and crossbred animals from Angus, Blond d'Aquitaine, Charolais, Gelbvieh, Hereford, Limousin, Maine-Anjou, Salers, Shorthorn, and Simmental breeds, estimated high correlations between maternal dominant and direct epistatic effects as well as between direct and maternal additive effects. In those cases, multicollinearity was responsible for an overestimation of variance components, a bias in estimates of genetic effects, and greater standard errors associated to regression coefficients. Consequently, the efficiency of selection and crossbreeding strategies based on these results was affected.

The objective of this work was to assess the degree of multicollinearity and to identify the variables involved in linear dependence relations in additive-dominant models.

Materials and Methods

Data of birth weight (BW), yearling weight (YW), and scrotal circumference (SC) from 149,469 animals of Montana Tropical breed born between 1994 and 2008 were used (Table 1). These individuals are progenies of 92,729 dams and 853 sires, providing genetic information from three generations (Brinks et al., 1961). The database is formed by animals reared in Brazil and Uruguay, and kept in tropical pastures, mostly in acid soils with *Urochloa* spp. grass.

Salt and mineral supplementation were given to the animals during all experimental period. Animals were grouped into contemporary groups (CG) that considered year of birth, herd, management group within herd, and sex. After initial selection, only animals with valid measurements and parentage information were kept in the database. Furthermore, records from the CG with less than five animals with valid measurements, with progenies of only one sire or formed by individuals with only one breed composition were deleted from the database.

Since Montana Tropical is a multibreed population, the individuals from the different breed compositions were grouped according to the NABC system (Ferraz et al., 1999; Mourão et al., 2007), in which breeds are classified into four biological types. The biological type N included *Bos indicus* breeds, such as Gyr, Guzerat, Indubrazil, Nellore, Tabapuan, Boran, and other Zebu breeds. The biological type A is characterized by *Bos taurus* cattle adapted to the tropics by natural or artificial selection, and descent of animals introduced by the colonizers, as, for example, Bonsmara and Belmont Red. The biological type B is formed by *Bos taurus* breeds with British origin, like Angus, Devon, and Hereford. The biological type C is typified by *Bos taurus* breeds from continental Europe, including Charolais, Limousin, and Simmental (Table 2).

Two models were considered. The first one, denominated RM, included the fixed effects of dam age class at calving: AOD_1 (less than 27 months of age), AOD_2 (between 27 and 41 months), AOD_3 (from 42 to 59 months), AOD_4 (between 60 and 119 months), AOD_5 (between 120 and 143 months), AOD_6 (from 144 to 167 months), and AOD_7 (more than 168 months). The covariates were associated to the direct (BTA, BTB, and BTC) and maternal (MBTA, MBTB, and MBTC) additive effects of the biological types and to the non-additive effects of direct (NxA, NxB, NxC, AxB, AxC, and BxC) and maternal (HM) heterozygosity. The second model, denominated R, considered the same effects of RM, with the exception of the maternal

Table 1. Number of observations, mean, standard deviation, and minimum and maximum values for birth weight, yearling weight, and scrotal circumference.

Trait	Observations	Mean	Standard deviation	Minimum	Maximum
Birth weight (kg)	141,567	32.38	4.090	23.00	42.00
Yearling weight (kg)	58,124	268.53	47.806	137.90	400.90
Scrotal circumference (cm)	20,371	28.00	3.90	17.00	39.00

additive effects. For scrotal circumference, the age of the animal at measurement was also included in the models.

Coefficients for direct (BTA, BTB, and BTC) and maternal additive effects (MBTA, MBTB, and MBTC) of biological types were equal to the proportion of each biological type in the breed composition of the calf and in the breed composition of the dam, respectively. Because the sum of the proportions of biological types is equal to one, direct and maternal additive effects of the biological type N were excluded from the statistical models. The same strategy was adopted for dam age class at calving. For this covariate, the fourth class (AOD_4) was also excluded.

The non-additive effects of heterozygosity were obtained by a linear relationship to the coefficients of direct heterozygosity (HD) and maternal total (HM), which were calculated by the following equations (Roso et al., 2005b),

$$H_D = 1 - \sum_{i=1}^{n=4} S_i \times D_i \quad \text{and} \quad H_M = 1 - \sum_{i=1}^{n=4} MGS_i \times MGD_i,$$

in which: the number 4 on top of the summation sign is the number of biological types (N, A, B, C); and S_i, D_i, MGS_i, and MGD_i are the fractions of the i^{th} biological type of sire, dam, grandsire, and granddam, respectively.

Multicollinearity diagnostics was based on the variance inflation factor (VIF) and on the study of the condition indexes (CI) and eigenvalues from the correlation matrix among explanatory variables, all obtained through the Proc Reg procedure from the statistical software SAS.

The variance inflation factor (VIF) for the predictor variable X_i was obtained by the equation $VIF_i = 1/(1 - R_i^2)$, in which: R_i^2 is the multiple determination coefficient for the linear regression of X_i on the other covariates. The VIF describes the increase in the coefficient variance in the presence of multicollinearity (Freund & Littell, 2000). Therefore, the VIF was used to distinguish which covariates are possibly involved in quasi-dependence relations. Generally, values greater than ten for the covariates in the model suggest the existence of multicollinearity as the cause of estimation problems, such as ambiguity in the identification of important predictor variables, direction and magnitude of regression coefficients contrary to the prior expectation or without biological

Table 2. Number of observations for birth weight, yearling weight, and scrotal circumference in each genetic group based on the NABC system.

Necessary condition	Obsevations		
	Birth weight	Yearling weight	Scrotal circumference
3/4 Genetic group			
60%< N <90%	1,901	616	114
60%< A <90%	47	16	3
60%< B <90%	2,836	843	311
60%< C <90%	187	35	4
Montana Tropical			
18.75< N <31.25% and 18.75< A <31.25% and 18.75< B <31.25% and 18.75< C <31.25%	10,766	5,795	2,709
18.75< N <31.25% and 18.75< A <31.25% and 43.75< B <56.25% and C <6.25%	2,891	1,574	667
18.75< N <31.25% and 43.75< A <56.25% and B< 6.25% and 18.75< C <31.25%	11,225	5,733	2,438
N <37.25% and 12.50< A <87.50% and B ≤75% and C ≤75% and (N+A) ≥25% and (B+C) ≤75%	50,804	25,440	11,660
Purebred			
N ≥ 90%	5,841	192	93
A ≥ 90%	294	61	9
B ≥ 90%	2,034	509	170
C ≥ 90%	2	-(1)	-
F_1			
40≤ N ≤60% and 40≤ A ≤60%	2,070	544	104
40≤ N ≤60% and 40≤ B ≤60%	38,558	12,880	1,367
40≤ N ≤60% and 40≤ C ≤60%	5,072	1,430	176
Other genetic groups(1)			
Every breed composition that does not comply with the conditions above	7,039	2,456	546

(1)No animal with valid measurements for this trait complies with the criteria of the respective genetic group.

significance, and unstable estimates of regression coefficients (Chatterjee & Hadi, 2006).

The determinant of the correlation matrix among the explanatory variables is equal to the product of the eigenvalues λ_i. In the presence of multicollinearity, these eigenvalues and, consequently, the determinant are small. The condition index is calculated as $CI = (\lambda_{max}/\lambda_i)^{0.5}$, in which λ_{max} is the largest eigenvalue and λ_i is the i^{th} eigenvalue of the correlation matrix. Therefore, high CI values are indicators of dependence between the covariates because λ_i will be close to zero. Based on this, the CI was used to determine the number of collinearities in the model. CI values between 10 and 30 indicate weak multicollinearity, whereas CI values greater than 30 suggest strong multicollinearity (Belsley, 1991).

To detect which covariates are involved in linear dependences, the decomposition of variance associated to the eigenvalues was carried out according to Belsley (1991): $Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X'X})^{-1} = \sigma^2\mathbf{V}\boldsymbol{\Lambda}^{-1}\mathbf{V'}$, in which: σ^2 is the estimated residual variance; $\mathbf{V}$ are the eigenvectors of the matrix; and $\boldsymbol{\Lambda}$ are the eigenvalues of the diagonal matrix. If $V = v_{ij}$ is the variance of the i^{th} element of $\boldsymbol{\beta}$, the variance of each parameter estimate can also be defined as the sum of the p components, with each number associated with an eigenvalue, as follows:

$$Var\left(\hat{\beta}_i\right) = \sigma^2 \sum_{j=1}^{p} \left(V_{ij}^2/\lambda_j\right),$$

in which p is the number of explanatory variables. Because the eigenvalues are in the denominator, the variance components associated with linear dependences (small λ_j) will be relatively high compared with the other components. Therefore, a high proportion of two or more coefficients related to small eigenvalues shows that the corresponding dependences are causing problems.

With $t_{ij} = v_{ij}^2/\lambda_j$ and $t_i = \sum_{j=i}^{p} t_{ij}$, the proportion of variance of the i^{th} regression coefficient associated with the j^{th} component of this decomposition will be obtained by the equation $\pi_{ij} = t_{ij}/t_i$, which i = 1, 2, ..., p. To detect multicollinearity, Belsley et al. (2004) recommend the identification of the eigenvalues with CI greater than 30. The variables with variance decomposition proportion (π_{ji}) greater than 0.5 for each of these eigenvalues are candidates to linear dependence.

Results and Discussion

Considering the RM model for birth weight (Figure 1 A), BTA, BTB, BTC, MBTB, MBTC, NxB, and NxC are probably involved in quasi-dependence relations, since they presented VIF greater than 10. Similarly, the covariates BTA, BTB, BTC, MBTB, MBTC, NxA, NxB, NxC, AxB, AxC, and BxC, for yearling weight (Figure 1 B), may be involved in multicollinearity, as well as BTA, BTB, BTC, MBTB, MBTC, NxA, NxB, NxC, AxC, and BxC for scrotal circumference (Figure 1 C).

With the reduction of covariates included in the analysis model (model R), a decrease was observed in the number of covariates involved in multicollinearity and in the VIF values. For birth weight, only BTC showed a VIF value greater than 10. For yearling weight, the covariates BTA, BTC, NxC, and AxC showed VIF values greater than the established threshold, whereas, for scrotal circumference, this was observed for the covariates BTC, NxC, AxC, and BxC. This result was already expected, since in the presence of multicollinearity the results are sensible to changes in the model and in the database (Belsley, 1991). Moreover, the exclusion of variables from the model is one alternative to mitigate multicollinearity effects on the results (Mason et al., 1975).

The same behavior was observed in the number of collinearities identified by CI values. For the RM model, two weak collinearities (CI=12.61 and 18.76) for birth weight, four weak collinearities (CI=10.00, 15.02, 19.79, and 26.77) for yearling weight, and three weak collinearities (CI=11.93, 18.33, and 24.73) for scrotal circumference were detected. Considering the R model, the following were observed: one weak collinearity (CI=11.28) for birth weight, two weak collinearities (CI=11.02 and 22.12) for yearling weight, and one weak collinearity (CI=21.17) for scrotal circumference. These CI values were associated with eigenvalues ranging from 0.01 to 0.03.

From the decomposition of the variance regression coefficients related to the largest condition index (CI=18.76) observed in the RM model for birth weight trait, BTB (π=0.69), MBTB (π=0.83), MBTC (π=0.68), and NxC (π=0.67) may have formed a linear dependence relation. Considering the second largest condition index (CI=12.61), the covariates BTC (π=0.55) and NxC (π=0.53) could be involved in a linear relationship. For the R model and CI equal to

11.28, the covariates C (π=0.97), NxC (π=0.90), AxC (π=0.70), and BxC (π=0.65) showed higher values than the threshold (π=0.5), which indicates a possible collinearity between these variables.

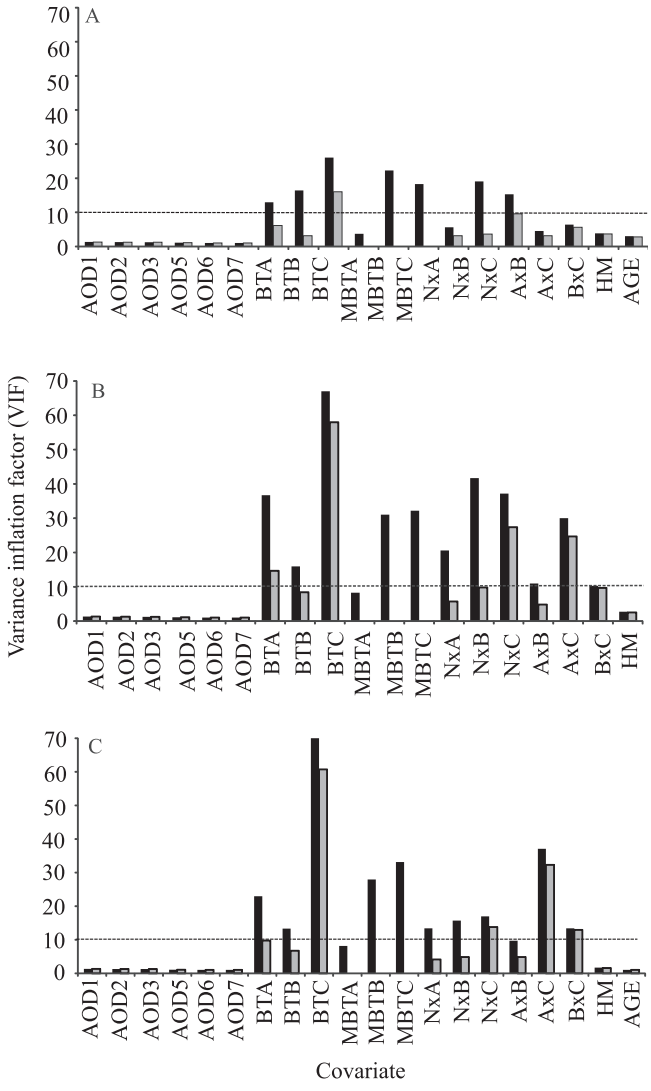

Figure 1. Variance inflation factor (VIF) for the explanatory covariates considered in the design matrix X of the models RM (■) and R (■) for birth weight (A), yearling weigth (B), and scrotal circumference (C). The dotted line (VIF>10) is an indicative of involvement in collinearity. AOD1 to AOD7 are the age classes of dam at calving; BTA, BTB, and BTC are the additive effects associated with the individual biological composition types for A, B, and C, respectively; MBTA, MBTB, and MBTC are the maternal additive effects associated with the maternal biological composition types for AM, BM, and CM, respectively; NxA, NxB, NxC, AxB, AxC, and BxC are the direct heterozygosity; and HM is the maternal total heterozygosity.

Similarly, for yearling weight with the RM model, the covariates BTC (π=0.71) and NxC (π=0.82), considering CI equal to 26.77, and BTA (π=0.81) and AxB (π=0.58), considering CI equal to 15.02, could be involved in linear relationships. For the collinearities identified by the condition indexes equal to 10.00 and 19.79, no variances with values above 0.5 were observed. However, for CI equal to 19.79, the covariates showed proportions of variance decomposition close to this threshold. Because this threshold value is empirical, the covariates MBTB (π=0.40), NxB (π=0.43), and AxC (π=0.46) could also be involved in multicollinearity. Considering the R model, the two weak collinearities identified by CI equal to 22.12 and 11.02 were represented by: BTC (π=0.98), NxC (π=0.88), AxC (π=0.92), and BxC (π=0.85); and BTA (π=0.79), BTB (π=0.64), and NxB (π=0.59), respectively.

The following covariates: BTC (π=0.90), NxC (π=0.83), AxC (π=0.59), and BxC (π=0.62); MBTB (π=0.62) and MBTC (π=0.56); and BTA (π=0.67) and AxB (π=0.61) were responsible for the collinearities identified by the condition indexes equal to 24.73, 18.33, and 11.93, respectively, obtained with the RM model for the scrotal circumference trait. Considering the R model, the single collinearity detected by the CI analyses, with a value of 21.17, was probably formed by the covariates BTC (π=0.98), NxC (π=0.84), AxC (π=0.94), and BxC (π=0.88).

The fixed effect of dam age class at calving (AOD) and the covariate of maternal heterozigoty (HM) were not involved in multicollinearity, since these variables were not detected in linear dependence relations nor by VIF or variance-decomposition proportions associated with the largest values of CI.

Multicollinearity can be a consequence of deficient sample data or of interrelationships among the variables that are inherent to the process under investigation (Chatterjee & Hadi, 2006). In these situations, not all combinations of predictor variables are represented by the data and, without data collected under all possible conditions, the effects of individual variables cannot be determined. Specifically for the present study, these circumstances occur due to the imbalance in the number of individuals in each genetic group, the restrictions imposed on breeding composition of the Montana Tropical breed (Table 2), and the natural connection between additive and non-additive effects,

since this is a crossbred population and the fraction of a biological type in the animal breed composition depends on the proportions from the other biological types. As an example, no purebred animals from biological type C were measured for yearling weight and scrotal circumference, and only two individuals from this genetic group were considered for birth weight analyses. Therefore, no performance information was observed in the sample when BTC was high and the fractions of the biological types A, B, and N were low. Similarly, it is not possible to obtain records when the fractions of biological types are all high or low, given that the sum of these proportions must be equal to one. As a result, independently of the diagnostic method used, the variables related to the additive and non-additive effects were detected as involved in a linear dependence relation, which was also reported by Rodríguez-Almeida et al. (1997) and Roso et al. (2005a, 2005b).

Based on multicollinearity causes, one solution to mitigate this regression problem is to collect additional data, so that more combinations among the explanatory variables are represented (Chatterjee & Hadi, 2006). In fact, an increase was observed in the degree of collinearity with the reduction in the amount of records by the comparison of VIF and CI values in the analyses for scrotal circumference and birth weight, which confirms the validity of this strategy. In this case, when multicollinearity is related to additive and non-additive effects, the additional data should involve representative individuals of several breed compositions and arrangements of biological types. However, it is often not possible to collect more data because of constraints on budget, time, and staff. Differences in requirements and production related to the diversity of animal size and growth rate make it difficult to maintain and sell cattle from diverse breed compositions in a same herd. Moreover, not all breeds can be used for beef production in Brazilian conditions of management and climate. Therefore, it is fairly difficult to ensure a balanced population for genetic analysis.

Another option to minimize multicollinearity complications in regression analysis is to reduce the number of covariates in the model. The means of VIF for the RM model were 8.66, 18.52, and 14.60 for birth weight, yearling weight, and scrotal circumference, respectively; whereas for the R model, these means were 3.99, 10.81, and 9.40, representing a reduction, in average, of 44% in the VIF mean due to the elimination of maternal additive effects from the model. However, this is not suitable, since these variables are important sources of variation for the traits. Furthermore, estimates of additive and non-additive genetic effects are useful for cross planning in the genetic improvement of traits of economic interest. Williams et al. (2010) reported direct and maternal breed effects for birth weight ranging from -0.5 to 10.1 kg and from -7.2 to 6.0 kg, respectively, as deviations from Angus breed. These authors also determined individual heterosis effects varying from 0.63 to -2.43 kg. Likewise, significant breed and heterosis effects were found by Perotto et al. (2000), Franke et al. (2001), Abdel-Aziz et al. (2003), Brandt et al. (2010), and Barichello et al. (2011) for birth weight, yearling weight, and scrotal circumference.

However, multicollinearity was not a severe problem in the models used in the present study, with variables only involved in weak collinearities. Similar results were observed by Roso et al. (2005a) in a crossbred population of 869,050 individuals formed by Angus, Blonde d'Aquitaine, Charolais, Gelbvieh, Hereford, Limousin, Maine-Anjou, Salers, Shorthorn, and Simmental breeds. For a model involving additive and dominance effects associated to these breeds, the authors estimated an average VIF of 26.81 and CI values between 3.41 and 38.85. In both studies, the amount of information available reduced multicollinearity effects, as mentioned before. Nevertheless, these effects are more evident in small samples. For example, Schoeman et al. (2002), employing 17,258 weaning weight records from a crossbred population formed by Afrikaner, Hereford, Angus, Simmental, and Charolais breeds, found VIF values from 1,386 to 19,402 for fixed direct and maternal additive effects, which are considerably larger than the threshold of 10.

Independently of the intensity, the presence of multicollinearity in regression models can affect the accuracy of the estimates and, consequently, the veracity of the inferences based on these results. Specifically in animal breeding programs, this can be translated in erroneous choices of breeds for crosses and in larger differences among the predicted and true genetic gain. Therefore, it is important to consider alternative approaches for the estimation of genetic effects when collinearity occurs. One of these methods is ridge regression (Hoerl & Kennard, 1970), which

is based on the addition of non-negative coefficients to the principal diagonal of the correlation matrix, which reduces or eliminates linear dependencies. Although biased, in the presence of multicollinearity, ridge estimators present lower standard errors and are more stable. Consequently, more accurate estimates are obtained than by using the ordinary least squares method. The least absolute shrinkage and selection operator (Lasso) regression (Tibshirani, 1996) is also a methodology used to deal with sparse solutions caused by multicollinearity, in which regression coefficients associated with irrelevant or redundant variables are reduced to zero. Roso et al. (2005a), Pimentel et al. (2006), Dias et al. (2011), Long et al. (2011), and Li & Sillanpää (2012) showed the advantages of these methodologies in regression models for genetic analyses when multicollinearity is present.

Conclusions

1. In the presence of multicollinearity, the results are sensitive to changes in the database and in the model.

2. Additive and non-additive effects are commonly involved in collinearity relations due to the inherent relationship between these variables.

3. Since the estimates yielded by the least squares method are less accurate when multicollinearity occurs, it is important to consider multicollinearity diagnostics as a preliminary analysis in animal breeding programs to avoid erroneous inferences and low selection efficiency.

4. It is still necessary to evaluate the impact of multicollinearity in the estimation of genetic parameters and breeding values and to assess the efficiency of an alternative method to the least squares methodology in order to complement the available information on this subject.

Acknowledgements

To Fundação de Amparo à Pesquisa do Estado de São Paulo (Fapesp), to Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and to Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes), for financial support; and to Grupo de Melhoramento Genético Animal e Biotecnologia from Universidade de São Paulo and to CFM-Leachman Pecuária Ltda., for support and for providing the database.

References

ABDEL-AZIZ, M.; SCHOEMAN, S.J.; JORDAAN, G.F. Estimation of additive, maternal and non-additive genetic effects of preweaning growth traits in a multibreed beef cattle project. **Animal Science Journal**, v.74, p.169-179, 2003.

BARICHELLO, F.; ALENCAR, M.M. de; TORRES JÚNIOR, R. de A.A.; SILVA, L.O.C. Environmental and genetic effects on weight, scrotal circumference and visual scores at weaning on Canchim beef cattle. **Revista Brasileira de Zootecnia**, v.40, p.286-293, 2011.

BELSLEY, D.A. **Conditioning diagnostics**: collinearity and weak data in regression. New York: John Wiley, 1991. 393p.

BELSLEY, D.A.; KUH, E.; WELSCH, R.E. **Regression diagnostics**: identifying influential data and sources of collinearity. New York: John Wiley, 2004. 292p.

BERGMANN, J.A.G.; HOHENBOKEN, W.D. Alternatives to least squares in multiple linear regression to predict production traits. **Journal of Animal Breeding and Genetics**, v.112, p.1-16, 1995.

BRANDT, H.; MÜLLENHOFF, A.; LAMBERTZ, C.; ERHARDT, G.; GAULY, M. Estimation of genetic and crossbreeding parameters for preweaning traits in German Angus and Simmental beef cattle and the reciprocal crosses. **Journal of Animal Science**, v.88, p.80-86, 2010.

BRINKS, J.S.; CLARK, R.T.; RICE, F.J. Estimation of genetic trends in beef cattle. **Journal of Animal Science**, v.20, p.903, 1961.

CASSADY, J.P.; YOUNG, L.D.; LEYMASTER, K.A. Heterosis and recombination effects on pig growth and carcass traits. **Journal of Animal Science**, v.80, p.2286-2302, 2002.

CHATTERJEE, S.; HADI, A.S. **Regression analysis by example**. 4.ed. New York: John Wiley, 2006. 408p.

DIAS, R.A.P.; PETRINI, J.; FERRAZ, J.B.S.; ELER, J.P.; BUENO, R.S.; COSTA, A.L.L. da; MOURÃO, G.B. Multicollinearity in genetic effects for weaning weight in a beef cattle composite population. **Livestock Science**, v.142, p.188-194, 2011.

FERRAZ, J.B.S.; ELER, J.P.; GOLDEN, B.L. Análise e genética do composto Montana Tropical. **Revista Brasileira de Reprodução Animal**, v.23, p.111-113, 1999.

FRANKE, D.E.; HABET, O.; TAWAH, L.C.; WILLIAMS, A.R.; DEROUEN, S.M. Direct and maternal genetic effects on birth and weaning traits in multibreed cattle data and predicted performance of breed crosses. **Journal of Animal Science**, v.79, p.1713-1722, 2001.

FREUND, R.J.; LITTELL, R.C. **SAS System for regression**. 3.ed. Cary: SAS Institute, 2000. 235p.

HOERL, A.E.; KENNARD, R.W. Ridge regression: biased estimation for nonorthogonal problems. **Technometrics**, v.12, p.55-68, 1970.

LI, Z.; SILLANPÄÄ, M.J. Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. **Theoretical and Applied Genetics**, v.125, p.419-435, 2012.

LONG, N.; GIANOLA, D.; ROSA, G.J.M.; WEIGEL, K.A. Dimension reduction and variable selection for genomic selection: application to predicting milk yield in Holsteins. **Journal of Animal Breeding and Genetics**, v.128, p.247-257, 2011.

LOPES, J.S.; RORATO, P.R.N.; WEBER, T.; ARAÚJO, R.O. de; DORNELLES, M. de A.; COMIN, J.G. Pre-weaning performance evaluation of a multibreed Aberdeen Angus x Nellore population using different genetic models. **Revista Brasileira de Zootecnia**, v.39, p.2418-2425, 2010.

MANSFIELD, E.R.; HELMS, B.P. Detecting multicollinearity. **The American Statistician**, v.36, p.158-160, 1982.

MASON, R.L.; GUNST, R.F.; WEBSTER, J.T. Regression analysis and problems of multicollinearity. **Communications in Statistics**, v.4, p.277-292, 1975.

MOURÃO, G.B.; FERRAZ, J.B.S.; ELER, J.P.; BALIEIRO, J.C.C.; BUENO, R.S.; MATTOS, E.C.; FIGUEIREDO, L.G.G. Genetic parameters for growth traits of a Brazilian *Bos taurus* x *Bos indicus* beef composite. **Genetics and Molecular Research**, v.6, p.1190-1200, 2007.

PEROTTO, D.; CUBAS, A.C.; MOLETTA, J.L.; LESSKIU, C. Heterosis upon weights in Canchim and Aberdeen Angus calves and in their reciprocal crosses. **Pesquisa Agropecuária Brasileira**, v.35, p.2511-2520, 2000.

PIMENTEL, E. da C.G.; QUEIROZ, S.A. de; CARVALHEIRO, R.; FRIES, L.A. Estimates of genetic effects in crossbred calves by different models and estimation methods. **Revista Brasileira de Zootecnia**, v.35, p.1020-1027, 2006.

RODRÍGUEZ-ALMEIDA, F.A.; VAN VLECK , L.D.; GREGORY, K.E. Estimation of direct and maternal breed effects for prediction of expected progeny differences for birth and weaning weights in three multibreed populations. **Journal of Animal Science**, v.75, p.1203-1212, 1997.

ROSO, V.M.; SCHENKEL, F.S.; MILLER, S.P.; SCHAEFFER, L.R. Estimation of genetic effects in the presence of multicollinearity in multibreed beef cattle evaluation. **Journal of Animal Science**, v.83, p.1788-1800, 2005a.

ROSO, V.M.; SCHENKEL, F.S.; MILLER, S.P.; WILTON, J.W. Additive, dominance, and epistatic loss effects on preweaning weight gain of crossbreed beef cattle from different *Bos taurus* breeds. **Journal of Animal Science**, v.83, p.1780-1787, 2005b.

SCHOEMAN, S.J.; AZIZ, M.A.; JORDAAN, G.F. The influence of multicollinearity on crossbreeding parameter estimates for weaning weight in beef cattle. **South African Journal of Animal Science**, v.32, p.239-246, 2002.

TIBSHIRANI, R. Regression shrinkage and selection via the Lasso. **Journal of the Royal Statistical Society Series B - Methodological**, v.58, p.267-288, 1996.

TORAL, F.L.B.; TORRES JÚNIOR, R.A. de A.; LOPES, P.S.; SILVA, L.O.C. da; REIS FILHO, J.C. Modeling the effect of the age of dam at calving on the weaning weight of Charolais-Zebu crossbred calves. **Revista Brasileira de Zootecnia**, v.38, p.1229-1237, 2009.

WILLIAMS, J.L.; AGUILAR, I.; REKAYA, R.; BERTRAND, J.K. Estimation of breed and heterosis effects for growth and carcass traits in cattle using published crossbreeding studies. **Journal of Animal Science**, v.88, p.460-466, 2010.

Received on November 26, 2011 and accepted on November 30, 2012