Brazilian Journal of Genetics, 19, 3, 479-484 (1996)
(Revista Brasileira de Genética)

METHODOLOGY

Diagnosis of multicollinearity: Assessment of the condition of correlation matrices used in genetic studies*

Samuel Pereira de Carvalho[1] *and Cosme Damião Cruz*[2]

ABSTRACT

Data from samples of bean plant genotypes, assessed in a trial carried out in the agricultural year 1991/92 in Viçosa, MG, were used to study multicollinearity. Different methods to diagnose the problem were applied to correlation matrices. The adverse consequences of multicollinearity become evident as it increases, because the variance associated with the estimates of the parameters increases simultaneously, making them less reliable. For the diagnosis of multicollinearity, the informal methods give only general guidance, and a combination of methods based on the estimation of the variance inflating factors and variance splitting allows not only quantification of the intensity of multicollinearity but also identification of the variables involved.

INTRODUCTION

Estimation of genetic parameters, based on multiple variables and using techniques involving correlation or covariance matrices such as path analysis and selection index, can be hampered by ill conditioning of the matrices caused by multicollinearity effects among the variables involved. Multicollinearity exists when the variables are correlated among themselves, but sometimes this term is applied only to cases where correlation among the variables is very high, or even perfect (Neter and Wasserman, 1974). The presence of multicollinearity in the working matrices can be diagnosed by techniques which provide information about its intensity and also identify the variables involved, allowing the adoption of measures to reduce possible adverse effects.

The adverse effects of multicollinearity have already been studied in relation to regression analysis. Hoerl and Kennard (1970) showed that multicollinearity can affect the square of the distance between the estimator of minimum squares $\hat{\beta}$ and the parameter β.

Webster *et al.* (1974) showed that the variance of $\hat{\beta}$ can be estimated by

$$V(\hat{\beta}_j) = (1 - R_j^2)^{-1} \sigma^2,$$

where R_j^2 is the multiple determination coefficient of the X_j regression over the remaining independent variables. If there is strong multicollinearity between X_j and some subset of the other independent variables, R_j^2 will

* Part of a thesis presented by S.P.C. to the Universidade Federal de Viçosa, in partial fulfillment of the requirements for the Doctoral degree.

[1] Departamento de Biologia, Universidade Federal de Lavras (UFLA), 37200-000 Lavras, MG, Brasil. Send correspondence to S.P.C.

[2] Departamento de Biologia Geral, Universidade Federal de Viçosa (UFV), 36570-000 Viçosa, MG, Brasil.

have a value near unity, making the variance of $\hat{\beta}_j$ very large.

According to Cruz and Regazzi (1994), multicollinearity can also affect the total determination coefficient ($\mathbf{R^2}$) of the path analysis. $\mathbf{R^2}$ corresponds to the determination of the base variable in function of the explicative variables in either the path diagram or the multiple linear regression. If the base variable is completely explained by the explicative variables, then $\mathbf{R^2 = 1}$, otherwise $\mathbf{R^2 < 1}$. However, determination coefficients larger than one are frequently found. Pereira (1984), for example, found $R^2 = 1.2823$ when studying the relationship between grain yield and its components in the common bean plant (*Phaseolus vulgaris*). The author did not analyze the possible causes for this high value, but it could have been caused by intense multicollinearity, as the variables considered, number of pods per plant, seed weight and number of seeds per pod are highly correlated.

Hayes and Hill (1981) explore the problem of selection indices which are calculated when multicollinearity is present. A method termed "bending" is proposed by these authors, for the modification of genetic and phenotypic covariance matrices.

An efficient procedure for the diagnosis of multicollinearity should directly reflect the intensity of its effects and allow the identification of the independent variables involved with this problem (Montgomery and Peck, 1981).

The aim of the present study was to use the methods for the diagnosis of multicollinearity quoted in the literature and to check the suitability of the correlation matrices to estimate the path coefficients.

MATERIAL AND METHODS

Samples from nine bean genotypes were assessed from seeds supplied by the Genetic Department at the Universidade Federal de Viçosa (UFV). Six paternal cultivars and five commercial cultivars, adapted to the climatic conditions of the region and used as check controls, were included in the experiment. The paternal cultivars belong to two groups, one a carrier of cold tolerance genes (Diacol Andino, Ica Tundama and Rojo 70) and other made of commercial cultivars adapted to the region where the experiment was carried out.

A randomized complete block design with four replications was used and each experimental plot was made up of two five-meter long rows, 0.60 m apart, totaling an area of 6.00 m^2. Fifteen seeds per linear meter of row were sown. A border of one of the cultivars adapted to the region was maintained around the experiment. Bean crop management practices recommended by Vieira (1983) were used. The following traits were assessed:

a) number of days to flowering (FLW);
b) number of days to maturity (MAT);
c) final number of plants per plot (PLP);
d) number of pods per plant (POD);
e) number of seeds per pod (SPD);
f) number of seeds per plant (SPL);
g) average weight of 100 seeds (AWS);
h) grain yield per plant (YPL);
i) straw weight (STW);
j) harvest index (HSI), given by the relationship between the grain weight and the sum of the grain and straw weights.

Additional information on the experimental data used in this work was presented by Carvalho (1995).

The following procedures were used to assess the presence of multicollinearity in the working matrices.

a) Test of the determinant of the correlation matrices

The matrix is in correlation form and, therefore, its determinant varies from zero to one. The determinant is one if the independent variables are orthogonal and zero if there is a complete linear dependence between them. As the determinant approaches zero multicollinearity becomes more intense. This method, however, does not allow the identification of the variables causing multicollinearity (Montgomery and Peck, 1981).

b) Analysis of the correlation matrix

This procedure involves the analysis of the non-diagonal elements (r_{ij}) of the correlation matrix. If the variables studied, $\mathbf{X_i}$ and $\mathbf{X_j}$ are approximately linear dependent, then r_{ij} will be near unity, in absolute values. A high correlation coefficient indicates multicollinearity. However, when the total number of independent variables is greater than two, this condition becomes only sufficient, but not necessary, and the absence of high correlation between two variables does not indicate absence of multicollinearity (Kmenta, 1971).

c) Analysis of eigenvalues and eigenvectors in the correlation matrix

When there are one or more close linear dependencies among the variables, one or more eigenvalues ($\lambda_1, \lambda_2, ..., \lambda_p$) of the correlation matrix will be small (Belsley *et al.*, 1980, and Silvey, 1969). Montgomery and Peck (1981) propose the assessment of the matrix condition number (**CN**), given its symmetry, defining **CN** as the relation between the largest and smallest eigenvalue. The authors point out that, if **CN < 100,** multicollinearity is not a serious problem. If **100 < CN < 1000,** multicollinearity is moderate to strong, and if **CN > 1000** there is severe multicollinearity.

The analysis of the eigenvalues can identify the approximate nature of the linear dependency existing between the variables (Belsley *et al.*, 1980). For these analyses, $R = V \Lambda V'$, where Λ is a diagonal matrix with dimensions **p x p**, (p is number of variables used to obtain the R correlation matrix), whose elements are the eigenvalues λ_j **(j = 1, 2, ..., p)** of R, and V is an orthogonal matrix with p x p dimension whose columns ($\mathbf{v_1}, \mathbf{v_2}, ..., \mathbf{v_p}$) are the normalized eigenvectors of R. An eigenvalue (λ_j) close to zero indicates linear dependence among the observations. The elements of the eigenvector ($\mathbf{v_j}$) associated with this eigenvalue describe the nature of this dependency.

d) Inflation factors of the variance

Marquardt (1970) gives the diagonal elements of the matrix $R_{xx}^{-1} = (X'X)^{-1}$ of variance inflation factors (VIF) when the matrix X'X is taken in the correlation form. These factors can be used to detect multicollinearity (Montgomery and Peck, 1981). The variance of the jth minimum squares regression coefficient is given by $\mathbf{v_{jj}}\,\sigma^2$, where $\mathbf{v_{jj}}$ can be interpreted as the factor that increases the variance of $\hat{\beta}$ when there is linear dependency among the variables. According to Neter *et al.* (1983), if VIF has values greater than 10, it is possible that the minimum squares regression coefficients associated with such values are highly affected by multicollinearity.

e) Splitting into singular values

Any matrix ${}_nX_p$, where n is observations and p is variables, can be broken down to X = UDV', according to Lawson and Hanson (1974) and Belsley *et al.* (1980). Then, $\mathbf{X'X = (UDV')'(UDV') = VD^2V' = V \Lambda V'}$, where **U** is a n x p dimension matrix, whose columns are the eigenvectors associated with the eigenvalues of X'X. Matrix V is of p x p dimension formed by the normalized eigenvectors of matrix X'X. The equality $\mathbf{U'U = V'V = I_p}$ exists. Matrix D is diagonal of p x p dimension, with non-negative diagonal elements μ_j (j = 1, 2, ..., p) representing the singular X values. Thus, X = UDV' is a form of splitting X into its singular values. According to Montgomery and Peck (1981), the ill conditioning degree of matrix X affects the size of the singular values, which will be larger as the singular value for each approximate linear dependency gets larger. Belsley *et al.* (1980) and Iemma (1988) called this relationship the condition index (η_k) of the X matrix, defining:

$$\eta_k = \frac{\mu_{max}}{\mu_k}, \quad k = 1, 2, ..., p.$$

Thus, $\eta_k \geq 1$, for all k.

The variance of the minimum squares estimator of β can be written as

$$V(\hat{\beta}) = \sigma^2 (X'X)^{-1} = \sigma^2 V \Lambda^{-1} V,$$

or for the kth component of $\hat{\beta}$,

$$V(\hat{\beta}_k) = \sigma^2 \sum_{j=1}^{p} \frac{v_{kj}^2}{\mu_j^2} = \sigma^2 \sum_{j=1}^{p} \frac{v_{kj}^2}{\lambda_j},$$

according to Ferrari (1989) and quoting Montgomery and Peck (1981) and Iemma (1987). Except for σ^2, the kth element of the $\mathbf{V \Lambda^{-1} V}$ diagonal is the kth inflation factor of the variance, thus

$$(VIF)_k = \sum_{j=1}^{p} \frac{v_{kj}^2}{\mu_j^2} = \sum_{j=1}^{p} \frac{v_{kj}^2}{\lambda_j}.$$

The presence of one or more small singular values, or eigenvalues, will cause inflation in the $\hat{\beta}_j$ variances.

The variance splitting process can also be done to measure the degree of multicollinearity, as Belsley *et al.* (1980) suggested:

$$\Pi_{jk} = \frac{\left(v_{kj}^2 / \mu_j^2 \right)}{(VIF)_k} \quad j = 1,2,...,p.$$

The elements Π_{jk} are ordered in a Π p x p dimension matrix. In this matrix, the elements of each column of Π are the proportions of the variance of each $\hat{\beta}_k$, which is also each variance inflation factor $(VIF)_K$ associated with the ith singular value.

High proportions of Π_{jk} associated with low μ_j values indicate that this singular value is associated with multicollinearity which, in turn, is inflating the $\hat{\beta}$ variances.

The following correlation matrices were used as working matrices to diagnose multicollinearity:

a) The A matrix - correlation matrix obtained from the log-transformed data of the primary components of grain yield, that is, the number of pods per plant (POD), number of seeds per pod (SPD) and average weight of 100 seeds (AWS), given by:

$$A = \begin{bmatrix} 1.0000 & 0.8288 & -0.7680 \\ 0.8288 & 1.0000 & -0.8723 \\ -0.7680 & -0.8723 & 1.0000 \end{bmatrix}$$

b) The B matrix - correlation matrix from the log-transformed data of the three primary components, but now also including the average number of seeds per plant (SPL). Since LOG(SPL) = LOG(POD) + LOG(SPD), the B matrix should show intense multicollinearity, resulting from the linear combination of the variables. Matrix B is given by:

$$B = \begin{bmatrix} 1.0000 & 0.9890 & 0.9023 & -0.8218 \\ 0.9890 & 1.0000 & 0.8288 & -0.7680 \\ 0.9023 & 0.8288 & 1.0000 & -0.8723 \\ -0.8218 & -0.7680 & -0.8723 & 1.0000 \end{bmatrix}$$

c) **The C matrix** - correlation matrix obtained from the secondary production components, such as the number of days to flowering (FLW), number of days to maturity (MAT), final number of plants per plot (PLP), straw weight (STW) and harvest index (HSI). The C matrix is given by:

$$C = \begin{bmatrix} 1.0000 & 0.2318 & -0.2488 & -0.5855 & -0.0919 \\ 0.2318 & 1.0000 & 0.0800 & 0.0689 & -0.8252 \\ 0.2488 & 0.0800 & 1.0000 & 0.4773 & -0.0393 \\ -0.5855 & 0.0689 & 0.4773 & 1.0000 & -0.0603 \\ -0.0919 & -0.8252 & -0.0393 & -0.0603 & 1.0000 \end{bmatrix}$$

RESULTS AND DISCUSSION

The methods were applied to diagnose multicollinearity in the working matrices using computer program GENES[3]. Tables I, II, and III show the results of these diagnoses referring to the matrices A, B, and C, respectively.

A method for diagnosing multicollinearity should supply information about the degree of manifestation and further, identify the variables involved in the problem. The method proposed by Neter *et al.* (1983) consists of a simple verification of the size of the non-diagonal elements of the matrix. Table I shows that, in matrix A, the largest correlation does not approach unity and, therefore, a high degree of multicollinearity is not expected in this case. The other methods must be applied to confirm this hypothesis.

The variables involved in simple correlations are identified by examining the correlation matrix. It is possible, however, that three or more independent variables are involved in multicollinearity without any pairs of these variables being highly correlated. A situation such as this is not diagnosed by the examination of correlation matrices, as pointed out by Ferrari (1989).

Table I shows that the variance inflation factors (VIF's) do not have a value higher than 10, also indicating the absence of a high degree of multicollinearity. In the correlation matrices the determinant varies from zero to one, if the variables are perfectly correlated or orthogonal among themselves, respectively. The determinant corresponding to the A matrix has no tendency to zero, in spite of being small, which is an indication of slight multicollinearity.

The eigenvalues and condition number tests, given by the ratio between the largest and smallest λ_j, identify the degree of multicollinearity present. Since the condition number found was less than 100, the multicollinearity can be classified as weak, according to the criteria of Montgomery and Peck (1981).

Table I - Results of the test of multicollinearity diagnosis applied to the A matrix.

Observed item	Results		
Largest correlation	0.8288 (LPOD x LSPD)		
Smallest correlation	-0.8723 (LSPD x LAWS)		
Number of VIF > 10	0		
Determinant	0.0728		
Eigenvalues	[2.6468	0.2372	0.1160]
Condition number	22.8155		
Singular value	[2.6468	0.2372	0.1160]
Condition index	[1.0000	11.1585	22.8172]
VIF_k	[4.4505	3.4469	5.3165]
Proportions resulting from	0.0103	0.0254	0.0018
splitting the direct effects	1.3839	0.1402	2.0890
variances	5.5442	7.7540	4.3079

LPOD: Logarithm of the average number of pods per plants in the plot; LSPD: logarithm of the average number of seeds per pod; LAWS: logarithm of the average weight of 100 seeds.

[3] Computer program created by the Genetic Department of Universidade Federal de Viçosa, 36570-000 Viçosa, MG, Brasil.

Table II - Results of the test of multicollinearity diagnosis applied to the B matrix.

Observed item	Results
Largest correlation	0.9890 (LPOD x LSPD)
Smallest correlation	-0.8723 (LAWS x LSPL)
Number of VIF > 10	3
Determinant	5.7141×10^{-6}
Eigenvalues	[3.59342 0.29052 0.11601 0.00005]
Condition number	76164.2800
Singular value	[3.59342 0.29052 0.11601 0.00005]
Condition index	[1.0000 12.3689 30.9751 71868.4000]
VIF_k	[4899.92 9800.59 6505.56 1.76]
Proportions resulting from splitting the direct effect variances	0.00 0.00 0.00 0.03 0.00 0.00 0.00 2.41 0.00 0.00 0.01 1.77 21182.07 21191.45 21177.19 2.19

LPOD: Logarithm of pod number per plant; LSPD: logarithm of seed number per pod; LAWS: logarithm of average weight of 100 seeds; LSPL: logarithm of number of seeds per plant.

Table III - Results of the multicollinearity diagnosis test applied to the C matrix.

Observed item	Results
Largest correlation	0.4773 (STW x PLP)
Smallest correlation	-0.8252 (MAT x HSI)
Number of VIF > 10	0
Determinant	0.1347
Eigenvalues	[1.91 1.85 0.73 0.35 0.15]
Condition number	12.9255
Singular value	[1.91 1.85 0.73 0.35 0.15]
Condition index	[1.00 1.03 2.62 5.46 12.73]
VIF_k	[2.23 2.96 1.37 2.30 3.18]
Proportions resulting from splitting the direct effect variances	0.03 0.01 0.06 0.03 0.01 0.03 0.03 0.00 0.00 0.05 0.09 0.10 0.84 0.09 0.00 0.76 0.60 0.09 1.89 0.08 3.71 4.43 1.57 1.41 5.94

MAT: Number of days to maturity; STW: straw weight; PLP: final number of plants per plot; HSI: harvest index.

These first methods do not provide much information about multicollinearity, because in spite of detecting its presence and even its intensity, they usually cannot identify the variables involved in the problem. The splitting of the correlation matrix into singular values and the determination of the condition index, given by the relationship between the largest and the other singular values, also give information on the degree of multicollinearity present. High values for the condition index and variance inflation factors (VIF_k) are indicators of high multicollinearity.

The combination of the methods, splitting the matrix into singular values with the splitting of the variance of the effects estimated for the path analysis, in analogy to the proposal of Belsley *et al.* (1980) for the analysis of multiple regression coefficients, allowed the identification of the variables involved in multicollinearity. In this way, the largest proportions of the variance corresponded to the variables 1 (logarithm of POD, LPOD) and 2 (logarithm of SPD, LSPD), which are the most inter-correlated variables.

Table II shows the results of tests to diagnose multicollinearity applied to the correlation matrix corresponding to four primary explicative variables (B matrix). One of these variables is made up of a combination of two others and must show intense multicollinearity. This problem can easily be detected by the near zero determinant and the excessively high condition number of the matrix. Also, the largest variance inflation factors of the estimators (VIF_k), associated with the largest variance component proportion, correspond to the variables LSPL (logarithm of SPL), LPOD and LSPD involved in a collinearity relationship. The high values found for VIF_k show the low degree of reliability which should be given to parameter estimates if the matrix is to be used in genetic studies as in path analysis or in estimation of selection indexes.

Table III shows the results of the tests to diagnose multicollinearity applied to the correlation matrix among the secondary explicative variables (C matrix). This matrix shows weak multicollinearity, with the condition number much smaller than 100. The VIF_k values and variance component proportions are also inferior to those presented in the other two examples. Thus, the parameter estimates obtained are associated with a small variance, and are, therefore, reliable.

RESUMO

Informações obtidas de amostras de materiais genotípicos de feijão, avaliados em ensaio conduzido em Viçosa, MG, durante o ano agrícola 1991/92, foram utilizadas para estudos de multicolinearidade. Matrizes de correlações foram submetidas a diferentes métodos para diagnóstico do problema. Constata-se que as conseqüências adversas da multicolinearidade tornam-se evidentes à medida em que esta aumenta, pois as variâncias associadas às estimativas dos

parâmetros aumentam simultaneamente, tornando essas estimativas menos confiáveis. Quanto ao diagnóstico da multicolinearidade observa-se que os métodos informais forneceram indicações apenas de caráter geral. Já a combinação dos métodos baseados na estimação dos fatores de inflação da variância e na decomposição desta possibilitaram não só a quantificação da intensidade com que a multicolinearidade se manifesta como também a identificação das variáveis envolvidas.

REFERENCES

Belsley, D.A., Kuh, E. and **Welch, R.E.** (1980). *Regression Diagnostics: Identifying Data and Sources of Collinearity*. John Wiley & Sons, New York, pp. 292.

Carvalho, S.P. (1995). Métodos alternativos de estimação de coeficientes de trilha e índices de seleção, sob multicolinearidade. Doctoral thesis, Universidade Federal de Viçosa, Viçosa, MG.

Cruz, C.D. and **Regazzi, A.J.** (1994). *Modelos Biométricos Aplicados ao Melhoramento Genético*. Universidade Federal de Viçosa, Viçosa, MG, pp. 390.

Ferrari, F. (1989). Estimadores viesados para modelos de regressão em presença de multicolinearidade. Doctoral thesis, Escola Superior de Agricultura "Luiz de Queiroz", USP, Piracicaba, SP.

Hayes, J.F. and **Hill, W.G.** (1981). Modification of estimates of parameters in the construction of the genetic selection indices ('Bending'). *Biometrics 37*: 483-493.

Hoerl, A.E. and **Kennard, R.W.** (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*: 55-68.

Iemma, A.F. (1987). Modelos lineares. Uma introdução para profissionais da pesquisa agropecuária. In: *Simpósio de Estatística Aplicada à Experimentação Agronômica. 32nd Reunião da Sociedade Brasileira de Biometria*, Londrina, PR, SBB, pp. 263.

Iemma, A.F. (1988). *Matrizes para Estatística*. Um texto para profissionais de ciências aplicadas. Escola Superior de Agricultura "Luiz de Queiroz", USP, Piracicaba, SP, pp. 339.

Kmenta, J. (1971). *Elements of Econometrics*. MacMillan Publishing Co., Inc., New York, pp. 655.

Lawson, C.L. and **Hanson, R.J.** (1974). *Solving Least Square Problems*. Prentice Hall, Englewood Cliffs, pp. 340.

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

Montgomery, D.C. and **Peck, E.A.** (1981). *Introduction to Linear Regression Analysis*. John Wiley & Sons, New York, pp. 504.

Neter, J. and **Wasserman, W.** (1974). *Applied Linear Statistical Models*. Richard D. Irwin Inc., Homewood, pp. 842.

Neter, J., Wasserman, W. and **Kutner, M.H.** (1983). *Applied Linear Regression Models*. Richard D. Irwin Inc., Homewood, pp. 547.

Pereira, T.N.S. (1984). Estimativas de parâmetros genéticos na identificação de progenitores para o melhoramento do feijoeiro-comum (*Phaseolus vulgaris* L.). Master's thesis, Universidade Federal de Viçosa, Viçosa, MG.

Silvey, S.D. (1969). Multicollinearity and imprecise estimation. *J. of the Royal Stat. Society*, Serie B. *31*: 539-552.

Vieira, C. (1983). *Cultura do Feijão*. 2nd edn. Universidade Federal de Viçosa, Viçosa, MG, pp. 146.

Webster, J.T., Gunst, R.F. and **Mason, R.L.** (1974). Latent root regression analysis. *Technometrics 16*: 513-522.

(Received January 12, 1995)